

Laurie A. Schintler
Connie L. McNeely
Editors

Encyclopedia of Big Data

Encyclopedia of Big Data

Laurie A. Schintler • Connie L. McNeely
Editors

Encyclopedia of Big Data

With 54 Figures and 29 Tables

 Springer

Editors

Laurie A. Schintler
George Mason University
Fairfax, VA, USA

Connie L. McNeely
George Mason University
Fairfax, VA, USA

ISBN 978-3-319-32009-0 ISBN 978-3-319-32010-6 (eBook)
ISBN 978-3-319-32011-3 (print and electronic bundle)
<https://doi.org/10.1007/978-3-319-32010-6>

© Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This encyclopedia was born in recognition of the fact that the big data revolution is upon us. Referring generally to data characterized by the “7 Vs” – volume, variety, velocity, variability, veracity, vulnerability, and value – big data is increasingly ubiquitous, impacting nearly every aspect of society in every corner of the globe. It has become an essential player and resource in today’s expanding digitalized and information-driven world and is embedded in a complex and dynamic ecosystem comprised of various industries, groups, algorithms, disciplines, platforms, applications, and enabling technologies.

On the one hand, big data is a critical driver of productivity, innovation, and well-being. In this context, various sources of big data – for example, satellites, digital sensors, observatories, crowdsourcing mechanisms, mobile devices, the World Wide Web, and the Internet of Things – coupled with advancing analytical and computational capacities and capabilities, continue to contribute to data-driven solutions and positive transformations across all sectors of the economy. On the other hand, the uses and applications of big data come with an array of challenges and problems, some of which are technical in nature and others that involve ethical, social, and legal dimensions that affect and are affected by societal constraints and considerations.

Understanding the opportunities and challenges brought on by the explosion of information that marks society today requires consideration of an encompassing array of questions and issues that arise with and because of big data. For example, the massive size and high dimensionality of big datasets present computational challenges and problems of validation linked not only to selection biases and measurement errors but also to spurious correlations and storage and scalability blockages.

Moreover, the bigger the data, the bigger the potential for its use and its misuse, whether relative to innovations and progress or referencing data asymmetries, ethical violations, discrimination, and biases. Accordingly, a wide range of topics, along with policies and strategies, regarding the nature and engagement of big data across levels of analysis, are needed to ensure that the possible benefits of big data are maximized while the downsides are minimized.

Against this backdrop, the Springer Nature *Encyclopedia of Big Data* offers a complex and diverse picture of big data viewed through a multidimensional technological and societal lens, considering related aspects and trends within and across different domains, disciplines, and sectors. Moreover, the field of

big data itself is highly fluid, with new analytical and processing modalities, concepts, and applications unfolding and evolving on an ongoing basis. Reflecting the breadth, depth, and dynamics of the field – and of the big data ecosystem itself – the *Encyclopedia of Big Data* is designed to provide a comprehensive, foundational, and cutting-edge perspective on the topic. It is intended to be a resource for various audiences – from the big data novice to the data scientist, from the researcher to the practitioner, and from the analyst to the generally interested lay public. The *Encyclopedia* has an international focus, covering the many aspects, uses, and applications of big data that transcend national boundaries. Accordingly, the *Encyclopedia of Big Data* draws upon the expertise and experience of leading scholars and practitioners from all over the world. Our aim is that it will serve as a valuable resource for understanding and keeping abreast of the constantly evolving, complex, and critical field of big data.

Fairfax, USA
January 2022

Laurie A. Schintler
Connie L. McNeely
Editors

List of Topics

Agile Data	Big Data Workforce
AgInformatics	Big Geo-data
Agriculture	Big Humanities Project
Algorithm	Big Variety Data
Algorithmic Complexity	Bioinformatics
American Bar Association	Biomedical Data
American Civil Liberties Union	Biometrics
American Library Association	Biosurveillance
Animals	Blockchain
Anomaly Detection	Blogs
Anonymity	Border Control/Immigration
Anonymization Techniques	Brand Monitoring
Anthropology	Business
Antiquities Trade, Illicit	Business Intelligence Analytics
Apple	Business-to-Community (B2C)
Archaeology	Cancer
Artificial Intelligence	Cell Phone Data
Arts	Census Bureau (U.S.)
Asian Americans Advancing Justice	Centers for Disease Control and Prevention (CDC)
Association Versus Causation	Charter of Fundamental Rights (EU)
Astronomy	Chemistry
Authoritarianism	Clickstream Analytics
Authorship Analysis and Attribution	Climate Change, Hurricanes/Typhoons/Cyclones
Automated Modeling/Decision Making	Climate Change, Rising Temperatures
Aviation	Cloud Computing
Behavioral Analytics	Cloud Services
Bibliometrics/Scientometrics	Collaborative Filtering
Big Data and Theory	Common Sense Media
Big Data Concept	Communication Quantity
Big Data Literacy	Communications
Big Data Quality	Complex Event Processing (CEP)
Big Data Research and Development Initiative (Federal, U.S.)	Complex Networks
Big Data Theory	Computational Social Sciences

Computer Science	Data Visualization
Content Management System (CMS)	Database Management Systems (DBMS)
Content Moderation	Datafication
Contexts	Data-Information-Knowledge-Action Model
Core Curriculum Issues (Big Data Research/ Analysis)	Data-Information-Knowledge-Wisdom (DIKW) Pyramid, Framework, Continuum
Corporate Social Responsibility	Decision Theory
Corpus Linguistics	Deep Learning
Correlation Versus Causation	De-identification/Re-identification
COVID-19 Pandemic	Demographic Data
Crowdsourcing	Digital Advertising Alliance
Cultural Analytics	Digital Divide
Curriculum, Higher Education, and Social Sciences	Digital Ecosystem
Curriculum, Higher Education, Humanities	Digital Knowledge Network Divide (DKND)
Cyber Espionage	Digital Literacy
Cyberinfrastructure (U.S.)	Digital Storytelling, Big Data Storytelling
Cybersecurity	Disaster Planning
Dashboard	Discovery Analytics, Discovery Informatics
Data Aggregation	Diversity
Data Architecture and Design	Driver Behavior Analytics
Data Brokers and Data Services	Drones
Data Center	Drug Enforcement Administration (DEA)
Data Cleansing	Earth Science
Data Discovery	E-Commerce
Data Exhaust	Economics
Data Fusion	Education and Training
Data Governance	Electronic Health Records (EHR)
Data Integration	Ensemble Methods
Data Integrity	Entertainment
Data Lake	Environment
Data Management and Artificial Intelligence (AI)	Epidemiology
Data Mining	Ethical and Legal Issues
Data Monetization	Ethics
Data Munging and Wrangling	European Commission
Data Processing	European Commission: Directorate-General for Justice (Data Protection Division)
Data Profiling	European Union
Data Provenance	European Union Data Protection Supervisor
Data Quality Management	Evidence-Based Medicine
Data Repository	Facebook
Data Science	Facial Recognition Technologies
Data Scientist	Financial Data and Trend Prediction
Data Sharing	Financial Services
Data Storage	Forestry
Data Streaming	Fourth Amendment
Data Synthesis	Fourth Industrial Revolution
Data Virtualization	Fourth Paradigm

France	Middle East
Gender and Sexuality	Mobile Analytics
Genealogy	Multiprocessing
Geography	Multi-threading
Google	National Association for the Advancement of
Google Analytics	Colored People
Google Books Ngrams	National Oceanic and Atmospheric
Google Flu	Administration
Governance	National Organization for Women
Granular Computing	National Security Agency (NSA)
Graph-Theoretic Computations/Graph Databases	Natural Hazards
Health Care Delivery	Natural Language Processing (NLP)
Health Informatics	Netflix
High Dimensional Data	Network Advertising Initiative
HIPAA	Network Analytics
Human Resources	Network Data
Humanities (Digital Humanities)	Neural Networks
Industrial and Commercial Bank of China	NoSQL (Not Structured Query Language)
Informatics	Nutrition
Information Commissioner, United Kingdom	Online Advertising
Information Overload	Online Identity
Information Quantity	Ontologies
Information Society	Open Data
Integrated Data System	Open-Source Software
Intelligent Transportation Systems (ITS)	Participatory Health and Big Data
Interactive Data Visualization	Patient Records
International Development	Patient-Centered (Personalized) Health
International Labor Organization	PatientsLikeMe
International Nongovernmental Organizations	Persistent Identifiers (PIDs) for Cultural Heritage
(INGOs)	Pharmaceutical Industry
Internet Association, The	Policy Analytics
Internet of Things (IoT)	Political Science
Internet: Language	Pollution, Air
Italy	Pollution, Land
Journalism	Pollution, Water
Keystroke Capture	Precision Population Health
Knowledge Management	Predictive Analytics
LexisNexis	Prevention
Link Prediction in Networks	Privacy
Link/Graph Mining	Probabilistic Matching
LinkedIn	Profiling
Machine Learning	Psychology
Maritime Transport	Recommender Systems
Mathematics	Regression
Media	Regulation
Medicaid	Religion
Metadata	Risk Analysis

R-Programming	Supercomputing, Exascale Computing, High
Salesforce	Performance Computing
Satellite Imagery/Remote Sensing	Supply Chain and Big Data
Scientometrics	Surface Web vs Deep Web vs Dark Web
Semantic/Content Analysis/Natural Language	Sustainability
Processing	Systems Science
Semiotics	Tableau Software
Semi-structured Data	Technological Singularity
Sensor Technologies	Telemedicine
Sentic Computing	Time Series Analytics
Sentiment Analysis	Transnational Crime
“Small” Data	Transparency
Smart Cities	Transportation Visualization
Social Media	Treatment
Social Media and Security	United Nations Educational, Scientific and
Social Network Analysis	Cultural Organization (UNESCO)
Social Sciences	Upturn
Socio-spatial Analytics	Visualization
South Korea	Voice User Interaction
Space Research Paradigm	Vulnerability
Spain	Web Scraping
Spatial Data	White House Big Data Initiative
Spatial Econometrics	White House BRAIN Initiative
Spatial Scientometrics	WikiLeaks
Spatiotemporal Analytics	Wikipedia
Standardization	World Bank
State Longitudinal Data System	Zappos
Storage	Zillow
Structured Query Language (SQL)	

About the Editors

Laurie A. Schintler George Mason University, Fairfax, VA, USA

Laurie A. Schintler, Ph.D., is an associate professor in the Schar School of Policy and Government at George Mason University, where she also serves as director for data and technology research initiatives in the Center for Regional Analysis. Dr. Schintler received her Ph.D. degree in regional and urban planning from the University of Illinois, Urbana-Champaign. Her primary areas of expertise and research lie at the intersection of big data, emerging technologies, complexity theory, regional development, information science, critical infrastructure, innovation, and policy analytics. A recent focal point of her research is on the determinants and impacts, and related challenges and opportunities of big data use in a regional and “smart city” context. She is also active in developing data-driven analytical methods for characterizing and modeling socio-spatial interaction and dynamics. Additionally, Dr. Schintler conducts research on the complex interplay between technological divides – including the big data divide – and related social disparities. Her research also addresses ethical and social impacts and other issues associated with the use of big data, artificial intelligence, blockchain – and emerging modes of human-machine interaction – in relation to policy and program development. Dr. Schintler is very professionally active, with numerous peer-reviewed publications, reports, conference proceedings, co-edited volumes, and grants and contracts.

Connie L. McNeely George Mason University, Fairfax, VA, USA

Connie L. McNeely, Ph.D., is a sociologist and professor in the Schar School of Policy and Government at George Mason University, where she is also the director of the Center for Science, Technology, and Innovation Policy. Her teaching and research address various aspects of science, technology, and innovation, big data, emerging technologies, public policy, and governance. Dr. McNeely has directed major projects on big data and digitalization processes, scientific networks, and broadening participation and inclusion in science and technology fields. Along with studies focused on applications of information technologies and informatics, she has conducted research concerning data democratization and data interoperability, leveraging large,

complex datasets to inform policy development and implementation. Her recent work has engaged related issues involving artificial intelligence and ethics, human-machine relations, digital divides, and big data and discovery analytics. She has ongoing projects examining institutional and cultural dynamics in matters of big data engagement and ethical and social impacts, with particular attention to questions of societal inequities and inequalities. Dr. McNeely has numerous publications and is active in several professional associations, serves as a reviewer and evaluator in a variety of programs and venues, and sits on several advisory boards and committees. Dr. McNeely earned her B.A. (A.B.) in sociology from the University of Pennsylvania and M.A. (A.M.) and Ph.D. in sociology from Stanford University.

Contributors

Natalia Abuín Vences Complutense University of Madrid, Madrid, Spain

Gagan Agrawal School of Computer and Cyber Sciences, Augusta University, Augusta, GA, USA

Nitin Agarwal University of Arkansas Little Rock, Little Rock, AR, USA

Rajeev Agrawal Information Technology Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, USA

Btihad Ajana King's College London, London, UK

Omar Alghushairy Department of Computer Science, University of Idaho, Moscow, ID, USA

Samer Al-khateeb Creighton University, Omaha, NE, USA

Gordon Alley-Young Department of Communications and Performing Arts, Kingsborough Community College, City University of New York, New York, NY, USA

Abdullah Alowairdhi Department of Computer Science, University of Idaho, Moscow, ID, USA

Rayan Alshamrani Department of Computer Science, University of Idaho, Moscow, ID, USA

Raed Alsini Department of Computer Science, University of Idaho, Moscow, ID, USA

Ashraf Althbati Department of Computer Science, University of Idaho, Moscow, ID, USA

Ines Amaral University of Minho, Braga, Minho, Portugal
Instituto Superior Miguel Torga, Coimbra, Portugal
Autonomous University of Lisbon, Lisbon, Portugal

Scott W. Ambler Disciplined Agile Consortium, Toronto, ON, Canada

R. Bruce Anderson Earth & Environment, Boston University, Boston, MA, USA

Florida Southern College, Lakeland, FL, USA

Janelle Applequist The Zimmerman School of Advertising and Mass Communications, University of South Florida, Tampa, FL, USA

Giuseppe Arbia Università Cattolica Del Sacro Cuore, Catholic University of the Sacred Heart, Rome, Italy

Claudia Arcidiacono Dipartimento di Agricoltura, Alimentazione e Ambiente, University of Catania, Catania, Italy

Lázaro M. Bacallao-Pino University of Zaragoza, Zaragoza, Spain
National Autonomous University of Mexico, Mexico City, Mexico

Jonathan Z. Bakdash Human Research and Engineering Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, USA

Paula K. Baldwin Department of Communication Studies, Western Oregon University, Monmouth, OR, USA

Warren Bareiss Department of Fine Arts and Communication Studies, University of South Carolina Upstate, Spartanburg, SC, USA

Feras A. Batarseh College of Science, George Mason University, Fairfax, VA, USA

Anamaria Berea Department of Computational and Data Sciences, George Mason University, Fairfax, VA, USA

Center for Complexity in Business, University of Maryland, College Park, MD, USA

Magdalena Bielenia-Grajewska Division of Maritime Economy, Department of Maritime Transport and Seaborne Trade, University of Gdansk, Gdansk, Poland

Intercultural Communication and Neurolinguistics Laboratory, Department of Translation Studies, University of Gdansk, Gdansk, Poland

Colin L. Bird Department of Chemistry, University of Southampton, Southampton, UK

Tobias Blanke Department of Digital Humanities, King's College London, London, UK

Camilla B. Bosanquet Schar School of Policy and Government, George Mason University, Arlington, VA, USA

Mustapha Bouakkaz University Amar Telidji Laghouat, Laghouat, Algeria

Jan Lauren Boyles Greenlee School of Journalism and Communication, Iowa State University, Ames, IA, USA

David Brown Southern New Hampshire University, University of Central Florida College of Medicine, Huntington Beach, CA, USA
University of Wyoming, Laramie, WY, USA

Stephen W. Brown Alliant International University, San Diego, CA, USA

Emilie Bruzelius Arnhold Institute for Global Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Department of Epidemiology, Joseph L. Mailman School of Public Health, Columbia University, New York, NY, USA

Kenneth Button Schar School of Policy and Government, George Mason University, Arlington, VA, USA

Erik Cambria School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

Steven J. Campbell University of South Carolina Lancaster, Lancaster, SC, USA

Pilar Carrera Universidad Carlos III de Madrid, Madrid, Spain

Daniel N. Cassenti U.S. Army Research Laboratory, Adelphi, MD, USA

Guido Cervone Geography, and Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, PA, USA

Wendy Chen George Mason University, Arlington, VA, USA

Yixin Chen Department of Communication Studies, Sam Houston State University, Huntsville, TX, USA

Tao Cheng SpaceTimeLab, University College London, London, UK

Yon Jung Choi Center for Science, Technology, and Innovation Policy, George Mason University, Fairfax, VA, USA

Davide Ciucci Università degli Studi di Milano-Bicocca, Milan, Italy

Deborah Elizabeth Cohen Smithsonian Center for Learning and Digital Access, Washington, DC, USA

Germán G. Creamer School of Business, Stevens Institute of Technology, Hoboken, NJ, USA

Francis Dalisay Communication & Fine Arts, College of Liberal Arts & Social Sciences, University of Guam, Mangilao, GU, USA

Andrea De Montis Department of Agricultural Sciences, University of Sassari, Sassari, Italy

Trevor Diehl Media Innovation Lab (MiLab), Department of Communication, University of Vienna, Wien, Austria

Dimitra Dimitrakopoulou School of Journalism and Mass Communication, Aristotle University of Thessaloniki, Thessaloniki, Greece

Derek Doran Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA

Patrick Doupe Arnhold Institute for Global Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Stuart Dunn Department of Digital Humanities, King's College London, London, UK

Ryan S. Eanes Department of Business Management, Washington College, Chestertown, MD, USA

Catherine Easton School of Law, Lancaster University, Bailrigg, UK

R. Elizabeth Griffin Dominion Astrophysical Observatory, British Columbia, Canada

Robert Faggian Centre for Regional and Rural Futures, Deakin University, Burwood, VIC, Australia

James H. Faghmous Arnhold Institute for Global Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Arash Jalal Zadeh Fard Department of Computer Science, University of Georgia, Athens, GA, USA

Vertica (Hewlett Packard Enterprise), Cambridge, MA, USA

Jennifer Ferreira Centre for Business in Society, Coventry University, Coventry, UK

Katherine Fink Department of Media, Communications, and Visual Arts, Pace University, Pleasantville, NY, USA

David Freet Eastern Kentucky University, Southern Illinois University, Edwardsville, IL, USA

Lisa M. Frehill Energetics Technology Center, Indian Head, MD, USA

Jeremy G. Frey Department of Chemistry, University of Southampton, Southampton, UK

Martin H. Frické University of Arizona, Tucson, AZ, USA

Kassandra Galvez Florida Southern College, Lakeland, FL, USA

Katherine R. Gamble U.S. Army Research Laboratory, Adelphi, MD, USA

Song Gao Department of Geography, University of California, Santa Barbara, CA, USA

Department of Geography, University of Wisconsin-Madison, Madison, WI, USA

Alberto Luis García Departamento de Ciencias de la Comunicación Aplicada, Facultad de Ciencias de la información, Universidad Complutense de Madrid, Madrid, Spain

Sandra Geisler Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

Matthew Geras Florida Southern College, Lakeland, FL, USA

Homero Gil de Zúñiga Media Innovation Lab (MiLab), Department of Communication, University of Vienna, Wien, Austria

Erik Goepner George Mason University, Arlington, VA, USA

Yessenia Gomez School of Public Health Institute for Applied Environmental Health, University of Maryland, College Park, MD, USA

Steven J. Gray The Bartlett Centre for Advanced Spatial Analysis, University College London, London, UK

Jong-On Hahm Department of Chemistry, Georgetown University, Washington, DC, USA

Rihan Hai RWTH Aachen University, Aachen, Germany

Muhiuddin Haider School of Public Health Institute for Applied Environmental Health, University of Maryland, College Park, MD, USA

Layla Hashemi Terrorism, Transnational Crime, and Corruption Center, George Mason University, Fairfax, VA, USA

James Haworth SpaceTimeLab, University College London, London, UK

Martin Hilbert Department of Communication, University of California, Davis, Davis, CA, USA

Kai Hoberg Kühne Logistics University, Hamburg, Germany

Mél Hogan Department of Communication, Media and Film, University of Calgary, Calgary, AB, Canada

Hemayet Hossain Centre for Regional and Rural Futures, Deakin University, Burwood, VIC, Australia

Gang Hua Visual Computing Group, Microsoft Research, Beijing, China

Fang Huang Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

Brigitte Huber Media Innovation Lab (MiLab), Department of Communication, University of Vienna, Wien, Austria

Carolynne Hultquist Geoinformatics and Earth Observation Laboratory, Department of Geography and Institute for CyberScience, The Pennsylvania State University, University Park, PA, USA

Suzi Iacono OIA, National Science Foundation, Alexandria, VA, USA

Ashiq Imran Department of Computer Science & Engineering, University of Texas at Arlington, Arlington, TX, USA

Ece Inan Girm American University Canterbury, Canterbury, UK

Elmira Jamei College of Engineering and Science, Victoria University, Melbourne, VIC, Australia

J. Jacob Jenkins California State University Channel Islands, Camarillo, CA, USA

Madeleine Johnson Centre for Regional and Rural Futures, Deakin University, Burwood, VIC, Australia

Patrick Juola Department of Mathematics and Computer Science, McAnulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Anirudh Kadadi Department of Computer Systems Technology, North Carolina A&T State University, Greensboro, NC, USA

Hina Kazmi George Mason University, Fairfax, VA, USA

Corey Koch Florida Southern College, Lakeland, FL, USA

Erik W. Kuiler George Mason University, Arlington, VA, USA

Joanna Kulesza Department of International Law and International Relations, University of Lodz, Lodz, Poland

Matthew J. Kushin Department of Communication, Shepherd University, Shepherdstown, WV, USA

Kim Lacey Saginaw Valley State University, University Center, MI, USA

Sabrina Lai Department of Civil and Environmental Engineering and Architecture, University of Cagliari, Cagliari, Italy

Paul Anthony Laux Lerner College of Business and Economics and J.P. Morgan Chase Fellow, Institute for Financial Services Analytics, University of Delaware, Newark, DE, USA

Simone Z. Leao City Futures Research Centre, Faculty of Built Environment, University of New South Wales, Sydney, NSW, Australia

Jooyeon Lee Hankuk University of Foreign Studies, Seoul, Korea (Republic of)

Joshua Lee Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Yulia A. Levites Strekalova College of Journalism and Communications, University of Florida, Gainesville, FL, USA

Loet Leydesdorff Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands

Meng-Hao Li George Mason University, Fairfax, VA, USA

Siona Listokin Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Kim Lorber Social Work Convening Group, Ramapo College of New Jersey, Mahwah, NJ, USA

Travis Loux Department of Epidemiology and Biostatistics, College for Public Health and Social Justice, Saint Louis University, St. Louis, MO, USA

Xiaogang Ma Department of Computer Science, University of Idaho, Moscow, ID, USA

Wolfgang Maass Saarland University, Saarbrücken, Germany

Marcienne Martin Laboratoire ORACLE [Observatoire Réunionnais des Arts, des Civilisations et des Littératures dans leur Environnement] Université de la Réunion Saint-Denis France, Montpellier, France

Lourdes S. Martinez School of Communication, San Diego State University, San Diego, CA, USA

Julian McAuley Computer Science Department, UCSD, San Diego, USA

Ernest L. McDuffie The Global McDuffie Group, Longwood, FL, USA

Ryan McGrady North Carolina State University, Raleigh, NC, USA

Heather McIntosh Mass Media, Minnesota State University, Mankato, MN, USA

Connie L. McNeely George Mason University, Fairfax, VA, USA

Esther Mead Department of Information Science, University of Arkansas Little Rock, Little Rock, AR, USA

John A. Miller Department of Computer Science, University of Georgia, Athens, GA, USA

Staša Milojević Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

Murad A. Mithani School of Business, Stevens Institute of Technology, Hoboken, NJ, USA

Giuseppe Modica Dipartimento di Agraria, Università degli Studi Mediterranea di Reggio Calabria, Reggio Calabria, Italy

David Cristian Morar Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Marco Morini Dipartimento di Comunicazione e Ricerca Sociale, Università degli Studi “La Sapienza”, Roma, Italy

Diana Nastasia Department of Applied Communication Studies, Southern Illinois University Edwardsville, Edwardsville, IL, USA

Sorin Nastasia Department of Applied Communication Studies, Southern Illinois University Edwardsville, Edwardsville, IL, USA

Alison N. Novak Department of Public Relations and Advertising, Rowan University, Glassboro, NJ, USA

Paul Nulty Centre for Research in Arts Social Science and Humanities, University of Cambridge, Cambridge, United Kingdom

Christopher Nyamful Department of Computer Systems Technology, North Carolina A&T State University, Greensboro, NC, USA

Daniel E. O’Leary Marshall School of Business, University of Southern California, Los Angeles, CA, USA

Barbara Cook Overton Communication Studies, Louisiana State University, Baton Rouge, LA, USA

Communication Studies, Southeastern Louisiana University, Hammond, LA, USA

Jeffrey Parsons Memorial University of Newfoundland, St. John's, Canada

Christopher Pettit City Futures Research Centre, Faculty of Built Environment, University of New South Wales, Sydney, NSW, Australia

William Pewen Department of Health, Nursing and Nutrition, University of the District of Columbia, Washington, DC, USA

Jürgen Pfeffer Bavarian School of Public Policy, Technical University of Munich, Munich, Germany

Matthew Pittman School of Journalism & Communication, University of Oregon, Eugene, OR, USA

Colin Porlezza IPMZ - Institute of Mass Communication and Media Research, University of Zurich, Zürich, Switzerland

Anirudh Prabhu Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

Sandeep Purao Bentley University, Waltham, USA

Christoph Quix Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

Hochschule Niederrhein University of Applied Sciences, Krefeld, Germany

Lakshmish Ramaswamy Department of Computer Science, University of Georgia, Athens, GA, USA

Ramón Reichert Department for Theatre, Film and Media Studies, Vienna University, Vienna, Austria

Sarah T. Roberts Department of Information Studies, University of California, Los Angeles, Los Angeles, CA, USA

Scott N. Romaniuk University of South Wales, Pontypridd, UK

Alirio Rosales University of British Columbia, Vancouver, Canada

Christopher Round George Mason University, Fairfax, VA, USA

Booz Allen Hamilton, Inc., McLean, VA, USA

Seref Sagioglu Department of Computer Engineering, Gazi University, Ankara, Turkey

Sergei A. Samoilenko George Mason University, Fairfax, VA, USA

Zerrin Savaşan Department of International Relations, Sub-Department of International Law, Selçuk University, Konya, Turkey

Deepak Saxena Indian Institute of Public Health Gandhinagar, Gujarat, India

Laurie A. Schintler George Mason University, Fairfax, VA, USA

Jon Schmid Georgia Institute of Technology, Atlanta, GA, USA

Hans C. Schmidt Pennsylvania State University – Brandywine, Philadelphia, PA, USA

Jason Schmitt Communication and Media, Clarkson University, Potsdam, NY, USA

Stephen T. Schroth Department of Early Childhood Education, Towson University, Baltimore, MD, USA

Raquel Vinader Segura Complutense University of Madrid, Madrid, Spain

Marc-David L. Seidel Sauder School of Business, University of British Columbia, Vancouver, BC, Canada

Kimberly F. Sellers Department of Mathematics and Statistics, Georgetown University, Washington, DC, USA

Padmanabhan Seshaiyer George Mason University, Fairfax, VA, USA

Alexander Sessums Florida Southern College, Lakeland, FL, USA

Mehdi Seyedmahmoudian School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, VIC, Australia

Salma Sharaf School of Public Health Institute for Applied Environmental Health, University of Maryland, College Park, MD, USA

Alan R. Shark Public Technology Institute, Washington, DC, USA
Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Kim Sheehan School of Journalism & Communication, University of Oregon, Eugene, OR, USA

Louise Shelley Terrorism, Transnational Crime, and Corruption Center, George Mason University, Fairfax, VA, USA

Marina Shilina Moscow State University (Russia), Moscow, Russia

Stephen D. Simon P. Mean Consulting, Leawood, KS, USA

Aram Sinnreich School of Communication, American University, Washington, DC, USA

Jörgen Skågeby Department of Media Studies, Stockholm University, Stockholm, Sweden

Christine Skubisz Department of Communication Studies, Emerson College, Boston, MA, USA

Department of Behavioral Health and Nutrition, University of Delaware, Newark, DE, USA

Mick Smith North Carolina A&T State University, Greensboro, NC, USA

Clare Southerton Centre for Social Research in Health and Social Policy Research Centre, UNSW, Sydney, Sydney, NSW, Australia

Ralf Spiller Macromedia University, Munich, Germany

Victor Sposito Centre for Regional and Rural Futures, Deakin University, Burwood, VIC, Australia

Alex Stojcevski School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, VIC, Australia

Veda C. Storey J Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA

Yulia A. Strekalova College of Journalism and Communications, University of Florida, Gainesville, FL, USA

Daniele C. Struppa Donald Bren Presidential Chair in Mathematics, Chapman University, Orange, CA, USA

Jennifer J. Summary-Smith Florida SouthWestern State College, Fort Myers, FL, USA

Culver-Stockton College, Canton, MO, USA

Melanie Swan New School University, New York, NY, USA

Yuzuru Tanaka Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan

Niccolò Tempini Department of Sociology, Philosophy and Anthropology and Egenis, Centre for the Study of the Life Sciences, University of Exeter, Exeter, UK

Doug Tewksbury Communication Studies Department, Niagara University, Niagara, NY, USA

Subash Thota Synectics for Management Decisions, Inc., Arlington, VA, USA

Ulrich Tiedau Centre for Digital Humanities, University College London, London, UK

Kristin M. Tolle University of Washington, eScience Institute, Redmond, WA, USA

Catalina L. Toma Communication Science, University of Wisconsin-Madison, Madison, WI, USA

Rochelle E. Tractenberg Collaborative for Research on Outcomes and – Metrics, Washington, DC, USA

Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and Rehabilitation Medicine, Georgetown University, Washington, DC, USA

Chiara Valentini Department of Management, Aarhus University, School of Business and Social Sciences, Aarhus, Denmark

Damien Van Puyvelde University of Glasgow, Glasgow, UK

Matthew S. VanDyke Department of Communication, Appalachian State University, Boone, NC, USA

Andreas Veglis School of Journalism and Mass Communication, Aristotle University of Thessaloniki, Thessaloniki, Greece

Natalia Abuín Vences Complutense University of Madrid, Madrid, Spain

Raquel Vinader Segura Complutense University of Madrid, Madrid, Spain
Rey Juan Carlos University, Fuenlabrada, Madrid, Spain

Jing Wang School of Communication and Information, Rutgers University, New Brunswick, NJ, USA

Anne L. Washington George Mason University, Fairfax, VA, USA

Nigel Waters Department of Geography and Civil Engineering, University of Calgary, Calgary, AB, Canada

Brian E. Weeks Communication Studies Department, University of Michigan, Ann Arbor, MI, USA

Adele Weiner Audrey Cohen School For Human Services and Education, Metropolitan College of New York, New York, NY, USA

Tao Wen Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA, USA

Carson C. Woo University of British Columbia, Vancouver, Canada

Rhonda Wrzenski Indiana University Southeast, New Albany, IN, USA

Masahiro Yamamoto Department of Communication, University at Albany – SUNY, Albany, NY, USA

Fan Yang Department of Communication Studies, University of Alabama at Birmingham, Birmingham, AL, USA

Qinghua Yang Department of Communication Studies, Texas Christian University, Fort Worth, TX, USA

Sandul Yasobant Center for Development Research (ZEF), University of Bonn, Bonn, Germany

Xinyue Ye Landscape Architecture & Urban Planning, Texas A&M University, College Station, TX, USA

Dzmitry Yuran School of Arts and Communication, Florida Institute of Technology, Melbourne, FL, USA

Ting Zhang Department of Accounting, Finance and Economics, Merrick School of Business, University of Baltimore, Baltimore, MD, USA

Weiwu Zhang College of Media and Communication, Texas Tech University, Lubbock, TX, USA

Bo Zhao College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA

Fen Zhao Alpha Edison, Los Angeles, CA, USA

A

Advanced Analytics

► Business Intelligence Analytics

Agile Data

Scott W. Ambler
Disciplined Agile Consortium, Toronto, ON,
Canada

To succeed at big data you must be able to process large volumes of data, data that is very often unstructured. More importantly, you must be able to swiftly react to emerging opportunities and insights before your competitor does. A Disciplined Agile approach to big data is evolutionary and collaborative in nature, leveraging proven strategies from the traditional, lean, and agile canons. Collaborative strategies increase both the velocity and quality of work performed while reducing overhead. Evolutionary strategies – those that deliver incremental value through iterative application of architecture and design modeling, database refactoring, automated regression testing, continuous integration (CI) of data assets, continuous deployment (CD) of data assets, and configuration management – build a solid data foundation that will stand the test of time. In effect this is the application of proven, leading-edge software engineering practices to big data.

This chapter is organized into the following sections:

1. Why Disciplined Agile Big Data?
2. Be Agile: An Agile Mindset for Data Professionals
3. Do Agile: The Agile Database Technique Stack
4. Last Words

Why Disciplined Agile Big Data?

The Big Data environment is complex. You are dealing with overwhelming amounts of data coming in from a large number of disparate data sources; the data is often of questionable quality and integrity, and the data is often coming from sources that are outside your scope of influence. You need to respond to quickly changing stakeholder needs without increasing the technical debt within your organization. It is clear that in the one extreme traditional approaches to data management are insufficiently responsive, yet at the other extreme, mainstream agile strategies (in particular Scrum) come up short for addressing your long-term data management ideas. You need a middle ground that combines techniques for just enough modeling and planning at the most responsible moments for doing so with engineering techniques that produce high-quality assets that are easily evolved yet will still stand

the test of time. That middle ground is Disciplined Agile Big Data.

Disciplined Agile (DA) (Ambler and Lines 2012) is a hybrid framework that combines strategies from a range of sources including Scrum, Agile Modeling, Agile Data, Unified Process, Kanban, traditional, and many other sources. DA promotes a pragmatic and flexible strategy for tailoring and evolving processes that reflect the situation that you face. A Disciplined Agile approach to Big Data leverages agile strategies architecture and design modeling and modern software engineering techniques. These practices, described below, are referred to as the agile database technique stack. The aim is to quickly meet the dynamic needs of the marketplace without short-changing the long-term viability of your organization.

Be Agile: An Agile Mindset for Data Professionals

In many ways agility is more of an attitude than a skillset. The common characteristics of agile professionals are:

- Willing to work closely with others, working in pairs or small teams as appropriate
- Pragmatic in that they are willing to do what needs to be done to the extent that it needs to be done
- Open minded, willing to experiment and learn new techniques
- Responsible and therefore willing to seek the help of the right person(s) for the task at hand
- Eager to work iteratively and incrementally, creating artifacts that are sufficient to the task at hand

Do Agile: The Agile Database Techniques Stack

Of course it isn't sufficient to "be agile" if you don't know how to "do agile." The following figure overviews the critical technical techniques required for agile database evolution. These agile

database techniques have been proven in practice and enjoy both commercial and open source tooling support (Fig. 1).

We say they form a stack because in order to be viable, each technique requires the one immediately below it. For it to make sense to continuously deploy database changes you need to be able to develop small and valuable vertical slices, which in turn require clean architecture and design, and so on. Let's explore each one in greater detail.

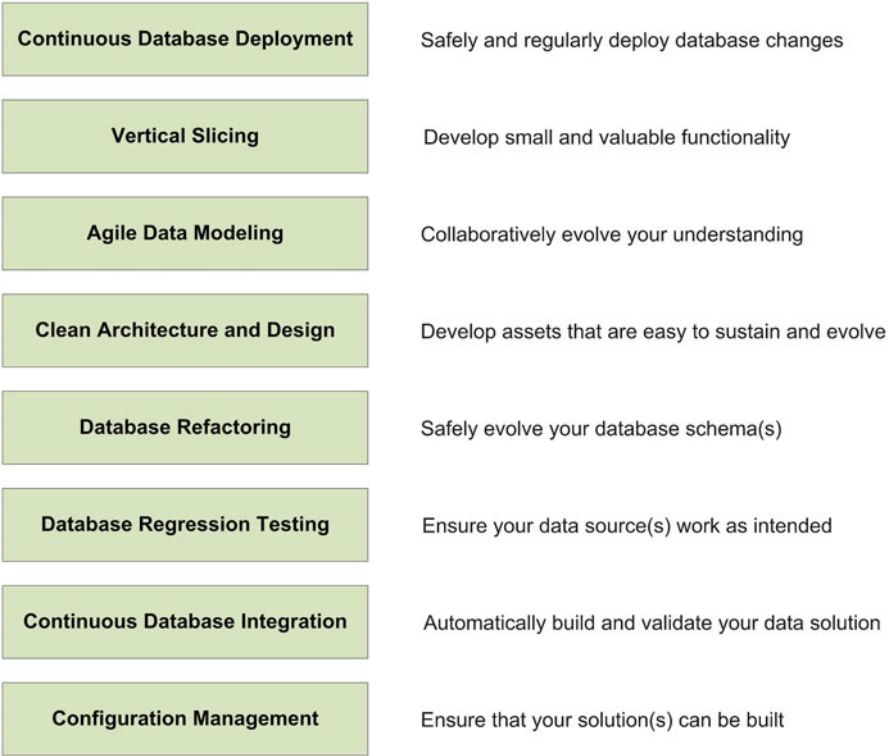
Continuous Database Deployment

Continuous deployment (CD) refers to the practice that when an integration build is successful (it compiles, passes all tests, and passes any automated analysis checks), your CD tool will automatically deploy to the next appropriate environment(s) (Sadalage 2003). This includes both changes to your business logic code as well as to your database. As you see in the following diagram, if the build runs successfully on a developer's work station their changes are propagated automatically into the team integration environment (which automatically invokes the integration build in that space). When the build is successful the changes are promoted into an integration testing environment, and so on (Fig. 2).

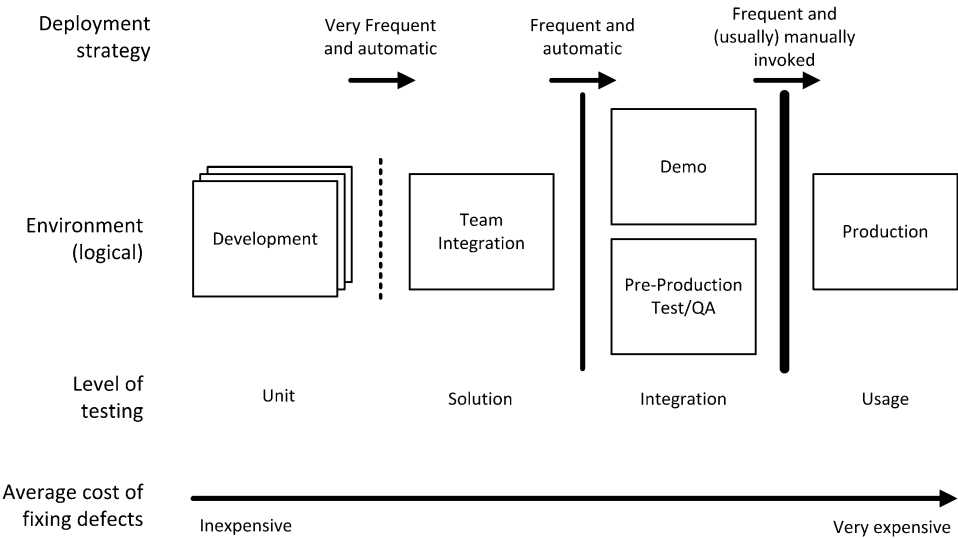
The aim of continuous database deployment is to reduce the time, cost, and risk of releasing database changes. Continuous database deployment only works if you are able to organize the functionality you are delivering into small, yet still valuable, vertical slices.

Vertical Slicing

A vertical slice is a top to bottom, fully implemented and tested piece of functionality that provides some form of business value to an end user. It should be possible to easily deploy a vertical slice into production upon request. A vertical slice can be very small, such as a single value on a report, the implementation of a business rule or calculation, or a new reporting view. For an agile team, all of this implementation work should be accomplished during a single iteration/sprint, typically a one- or two-week period. For teams following a lean delivery lifecycle, this timeframe



Agile Data, Fig. 1 The agile database technique stack



Agile Data, Fig. 2 Continuous database deployment

typically shrinks to days and even hours in some cases.

For a Big Data solution, a vertical slice is fully implemented from the appropriate data sources all the way through to a data warehouse (DW), data mart (DM), or business intelligence (BI) solution. For the data elements required by the vertical slice, you need to fully implement the following:

- Extraction from the data source(s)
- Staging of the raw source data (if you stage data)
- Transformation/cleansing of the source data
- Loading the data into the DW
- Loading into your data marts (DMs)
- Updating the appropriate BI views/reports where needed

A key concept is that you only do the work for the vertical slice that you're currently working on. This is what enables you to get the work done in a matter of days (and even hours once you get good at it) instead of weeks or months. It should be clear that vertical slicing is only viable when you are able to take an agile approach to modeling.

Agile Data Modeling

Many traditional data professionals believe that they need to perform detailed, up-front requirements, architecture, and design modeling before they can begin construction work. Not only has this been shown to be an ineffective strategy in general, when it comes to the dynamically evolving world of Big Data environments, it also proves to be disastrous. A Disciplined Agile approach strives to keep the benefits of modeling and planning, which are to think things through, yet avoid the disadvantages associated with detailed documentation and making important decisions long before you need to. DA does this by applying light-weight *Agile Modeling* (Ambler 2002) strategies such as:

1. **Initial requirements envisioning.** This includes both usage modeling, likely via user stores and epics, and conceptual modeling. These models are high-level at first; their

details will be fleshed out later as construction progresses.

2. **Initial architecture envisioning.** Your architecture strategy is typically captured in a free-form architecture diagram, network diagram, or UML deployment diagram. Your model(s) should capture potential data sources; how data will flow from the data sources to the target data warehouse(s) or data marts; and how that work flows through combinations of data extraction, data transformation, and data loading capabilities.
3. **Look-ahead modeling.** Sometimes referred to as “backlog refinement” or “backlog grooming,” the goal of look-ahead modeling is to explore work that is a few weeks in the future. This is particularly needed in complex domains where there may be a few weeks of detailed data analysis required to work through the semantics of your source data. For teams taking a sprint/iteration-based approach, this may mean that during the current iteration someone(s) on the team explores requirements to be implemented one or two iterations in the future.
4. **Model storming.** This is a just-in-time (JIT) modeling strategy where you explore something through in greater detail, perhaps work through the details of what a report should look like or how the logic of a business calculation should work.
5. **Test-driven development (TDD).** With TDD, your tests both validate your work and specify it. Specification can be done at the requirements level with acceptance tests and at the design level with developer tests. More on this later.

Clean Architecture and Design

High-quality IT assets are easier to understand, to work with, and to evolve. In many ways, clean architecture and design are fundamental enablers of agility in general. Here are a few important considerations for you:

1. **Choose a data warehouse architecture paradigm.** Although there is something to be said about both the Inmon and Kimball strategies, I

generally prefer DataVault 2 (Lindstedt and Olschimke 2015). DataVault 2 (DV2) has its roots in the Inmon approach, bringing learnings in from Kimball and more importantly practical experiences dealing with DW/BI and Big Data in a range of situations.

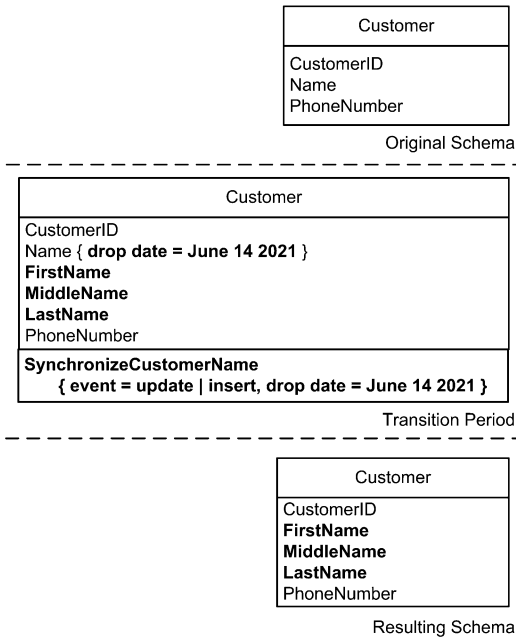
- 2. **Focus on loose coupling and high cohesion.** When a system is loosely coupled, it should be easy to evolve its components without significant effects on other components. Components that are highly cohesive do one thing and one thing only, in data parlance they are “highly normalized.”
- 3. **Adopt common conventions.** Guidelines around data naming conventions, architectural guidelines, coding conventions, user experience (UX) conventions, and others promote greater consistency in the work produced.
- 4. **Train and coach your people.** Unfortunately few IT professionals these days get explicit training in architecture and design strategies, resulting in poor quality work that increases your organization’s overall technical debt.

Database Refactoring

A refactoring is a simple change to your design that improves its quality without changing its semantics in a practical manner. A database refactoring is a simple change to a database schema that improves the quality of its design OR improves the quality of the data that it contains (Ambler and Sadalage 2006). Database refactoring enables you to safely and easily evolve database schemas, including production database schemas, over time by breaking large changes into a collection of smaller less-risky changes. Refactoring enables you to keep existing clean designs of high quality and to safely address problems in poor quality implementations.

Let’s work through an example. The following diagram depicts three stages in the life of the Split Column database refactoring. The first stage shows the original database schema where we see that the Customer table has a Name column where the full name of a person is stored. We have decided that we want to improve the quality of this table by splitting the column into three – in this case FirstName, MiddleName, and LastName.

The second stage, the transition period, shows how Customer contains both the original version of the schema (the Name column), the new/ desired version of the schema, and scaffolding code to keep the two columns in sync. The transition period is required so as to give the people responsibility for any systems that access customer name to update their code to instead work with the new columns. This approach is based on the Java Development Kit (JDK) deprecation strategy. The scaffolding code, in this case a trigger that keeps the four columns consistent with one another, is required so that the database maintains integrity over the transition period. There may be hundreds of systems accessing this information – at first they will all be accessing the original schema but over time they will be updated to access the new version of the schema – and because these systems can not all be reworked at once the database must be responsible for its own integrity. Once the transition period ends and the existing systems that access the Customer table have been update accordingly, the original schema and the scaffolding code can be removed safely (Fig. 3).



Agile Data, Fig. 3 Example database refactoring

Automated Database Testing

Quality is paramount for agility. Disciplined Agile teams will develop, in an evolutionary manner of course, an automated regression test suite that validates their work. They will run this test suite many times a day so as to detect any problems as early as possible. Automated regression testing like this enables teams to safely make changes, such as refactorings, because if they inject a problem they will be able to quickly find and then fix it.

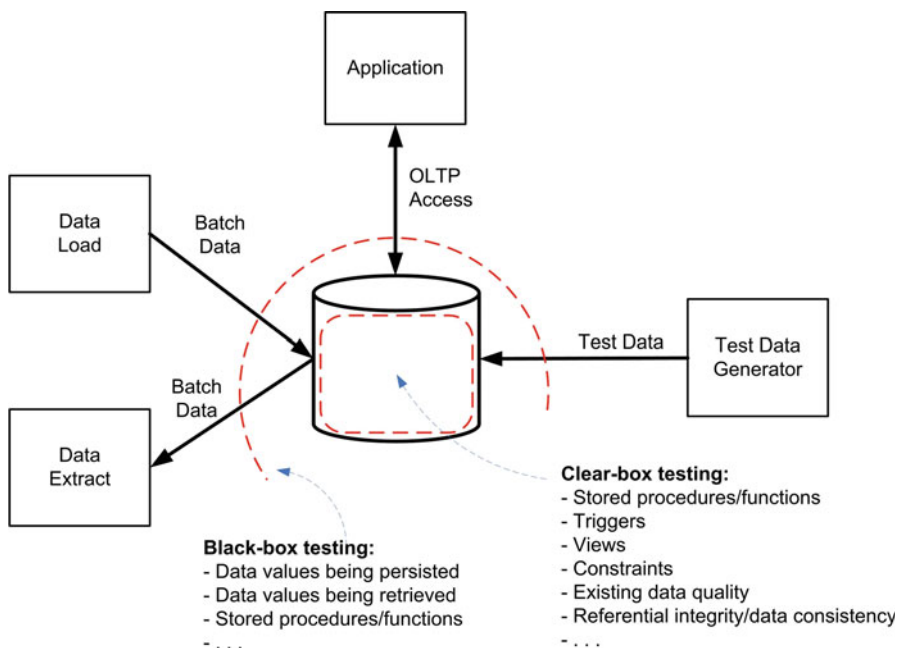
When it comes to testing, a database the following diagram summarizes the kind of tests that you should consider implementing (Ambler 2013). Of course there is more to testing Big Data implementations than this, you will also want to develop automated tests/checks for the entire chain from data sources through your data processing architecture into your DW/BI solution (Fig. 4).

In fact, very disciplined teams will take a test-driven development (TDD) approach where they write tests before they do the work to implement the functionality that the tests validate (Guernsey 2013). As a result, the tests do double duty – they

both validate and specify. You can do this at the requirements level by writing user acceptance tests, a strategy referred to as behavior driven design (BDD) or acceptance test driven design (ATDD), and at the design level via developer tests. By rethinking the order in which you work, in this case by testing first not last, you can streamline your approach while you increase its quality.

Continuous Database Integration

Continuous integration (CI) is a technique where you automatically build and test your system every time someone checks in a code change (Sadalage 2003). Disciplined agile developers will typically update a few lines of code, or make a small change to a configuration file, or make a small change to a PDM and then check their work into their configuration management tool. The CI tool monitors this, and when it detects a check, it automatically kicks off the build and regression test suite in the background. This provides very quick feedback to team members, enabling them to detect issues early.



Agile Data, Fig. 4 What to test in a database

Configuration Management

Configuration management is at the bottom of the stack, providing a foundation for all other agile database techniques. In this case there is nothing special about the assets that you are creating – ETL code, configuration files, data models, test data, stored procedures, and so on – in that if they are worth creating then they are also worth putting under CM control.

Last Words

I would like to end with two simple messages: First, you can do this. Everything described in this chapter is pragmatic, supported by tooling, and has been proven in practice in numerous contexts. Second, you need to do this. The modern, dynamic business environment requires you to work in a reactive manner that does not short change your organization's future. The Disciplined Agile approach described in this chapter describes how to do exactly that.

Further Reading

- Ambler, S. W. (2002). *Agile modeling: Effective practices for extreme programming and the unified process*. New York: Wiley.
- Ambler, S. W. (2013). *Database testing: How to regression test a relational database*. Retrieved from <http://www.agiledata.org/essays/databaseTesting.html>.
- Ambler, S. W., & Lines, M. (2012). *Disciplined agile delivery: A practitioner's guide to agile software delivery in the enterprise*. New York: IBM Press.
- Ambler, S. W., & Sadalage, P. J. (2006). *Refactoring databases: Evolutionary database design*. Boston: Addison Wesley.
- Guernsey, M., III. (2013). *Test-driven database development: Unlocking agility*. Upper Saddle River: Addison-Wesley Professional.
- Lindstedt, D., & Olschimke, M. (2015). *Building a scalable data warehouse with database 2.0*. Waltham: Morgan Kaufman.
- Sadalage, P. J. (2003). *Recipes for continuous database integration: Evolutionary database development*. Upper Saddle River: Addison-Wesley Professional.

AgInformatics

Andrea De Montis¹, Giuseppe Modica² and Claudia Arcidiacono³

¹Department of Agricultural Sciences, University of Sassari, Sassari, Italy

²Dipartimento di Agraria, Università degli Studi Mediterranea di Reggio Calabria, Reggio Calabria, Italy

³Dipartimento di Agricoltura, Alimentazione e Ambiente, University of Catania, Catania, Italy

Synonyms

E-agriculture; Precision agriculture; Precision farming

Definition

The term stems from the blending of the two words agriculture and informatics and refers to the application of informatics to the analysis, design and development of agricultural activities. It overarches expressions such as Precision Agriculture (PA), Precision Livestock Farming (PLF), and Agricultural landscape analysis and planning. The adoption of AgInformatics can accelerate agricultural development by providing farmers and decision makers with more accessible, complete, timely, and accurate information. However, it is still hindered by a number of important yet unresolved issues including big data handling, multiple data sources and limited standardization, data protection, and lack of optimization models. Development of knowledge-based systems in the farming sector would require key components, supported by Internet of things (IoT), data acquisition systems, ubiquitous computing and networking, machine-to-machine (M2M) communications, effective management of geospatial and temporal data, and ICT-supported cooperation among stakeholders.

Generalities

This relatively new expression derives from a combination of the two terms agriculture and informatics, hence alluding to the application of informatics to the analysis, design, and development of agricultural activities. It broadly involves the study and practice of creating, collecting, storing and retrieving, manipulating, classifying, and sharing information concerning both natural and engineered agricultural systems. The domains of application are mainly agri-food and environmental sciences and technologies, while sectors include biosystems engineering, farm management, crop production, and environmental monitoring. In this respect, it encompasses the management of the information coming from applications and advances of information and communication technologies (ICTs) in agriculture (e.g., global navigation satellite system, GNSS; remote sensing, RS; wireless sensor networks, WSN; and radio-frequency identification, RFID) and performed through specific agriculture information systems, models, and methodologies (e.g., farm management information systems, FMIS; GIScience analyses; Data Mining; decision support systems, DSS).

AgInformatics is an umbrella concept that includes and overlaps issues covered in precision agriculture (PA), precision livestock farming (PLF), and agricultural landscape analysis and planning, as follows.

Precision Agriculture (PA)

PA was coined in 1929 and later defined as “a management strategy that uses information technologies to bring data from multiple sources to bear on decisions associated with crop production” (Li and Chung 2015). The concept evolved since the late 1980s due to new fertilization equipment, dynamic sensing, crop yield monitoring technologies, and GNSS technology for automated machinery guidance.

Therefore, PA technology has provided farmers with the tools (e.g., built-in sensors in farming machinery, GIS tools for yield monitoring and mapping, WSNs, satellite and low-

altitude RS by means of unmanned aerial systems (UAS), and recently robots) and information (e.g., weather, environment, soil, crop, and production data) needed to optimize and customize the timing, amount, and placement of inputs including seeds, fertilizers, pesticides, and irrigation, activities that were later applied also inside closed environments, buildings, and facilities, such as for protected cultivation.

To accomplish the operational functions of a complex farm, FMISs for PA are designed to manage information about processes, resources (materials, information, and services), procedures and standards, and characteristics of the final products (Sørensen et al. 2010). Nowadays dedicated FMISs operate on networked online frameworks and are able to process a huge amount of data. The execution of their functions implies the adoption of various management systems, databases, software architectures, and decision models. Relevant examples of information management between different actors are supply chain information systems (SCIS) including those specifically designed for traceability and supply chain planning.

Recently, PA has evolved to predictive and prescriptive agriculture. Predictive agriculture regards the activity of combining and using a large amount of data to improve knowledge and predict trends, whereas prescriptive agriculture involves the use of detailed, site-specific recommendations for a farm field. Today PA embraces new terms such as precision citrus farming, precision horticulture, precision viticulture, precision livestock farming, and precision aquaculture (Li and Chung 2015).

Precision Livestock Farming (PLF)

The increase in activities related to livestock farming triggered the definition of the new term precision livestock farming (PLF), namely, the real-time monitoring technologies aimed at managing the smallest manageable production unit's temporal variability, known as “the per animal approach” (Berckmans 2004). PLF consists in the real-time gathering of data related to livestock animals and their close environment, applying

knowledge-based computer models, and extracting useful information for automatic monitoring and control purposes. It implies monitoring animal health, welfare, behavior, and performance and the early detection of illness or a specific physiological status and unfolds in several activities including real-time analysis of sounds, images, and accelerometer data, live weight assessment, condition scoring, and online milk analysis. In PLF, continuous measurements and a reliable prediction of variation in animal data or animal response to environmental changes are integrated in the definition of models and algorithms that allow for taking control actions (e.g., climate control, feeding strategies, and therapeutic decisions).

Agricultural Landscape Analysis and Planning

Agricultural landscape analysis and planning is increasingly based on the development of interoperable spatial data infrastructures (SDIs) that integrate heterogeneous multi-temporal spatial datasets and time-series information.

Nearly all agricultural data has some form of spatial component, and GISs allow to visualize information that might otherwise be difficult to interpret (Pierce and Clay 2007).

Land use/land cover (LU/LC) change detection methods are widespread in several research fields and represent an important issue dealing with the modification analysis of agricultural uses. In this framework, RS imagery plays a key role and involves several steps dealing with the classification of continuous radiometric information remotely surveyed into tangible information, often exposed as thematic maps in GIS environments, and that can be utilized in conjunction with other data sets. Among classification techniques, object-based image analysis (OBIA) is one of the most powerful techniques and gained popularity since the early 2000s in extracting meaningful objects from high-resolution RS imagery.

Proprietary data sources are integrated with social data created by citizens, i.e., volunteered geographic information (VGI). VGI includes crowdsourced geotagged information from social

networks (often provided by means of smart applications) and geospatial information on the Web (GeoWeb). Spatial decision support systems (SDSSs) are computer-based systems that help decision makers in the solution of complex problems, such as in agriculture, land use allocation, and management. SDSSs implement diverse forms of multi-criteria decision analysis (MCDA). GIS-based MCDA can be considered as a class of SDSS. Implementing GIS-MCDA within the World Wide Web environment can help to bridge the gap between the public and experts and favor public participation.

Conclusion

Technologies have the potential to change modes of producing agri-food and livestock. ICTs can accelerate agricultural development by providing more accessible, complete, timely, or accurate information at the appropriate moment to decision makers. Concurrently, management concepts, such as PA and PLF, may play an important role in driving and accelerating adoption of ICT technologies. However, the application of PA solutions has been slow due to a number of important yet unresolved issues including big data handling, limited standardization, data protection, and lack of optimization models and depends as well on infrastructural conditions such as availability of broadband internet in rural areas. The adoption of FMISs in agriculture is hindered by barriers connected to poor interfacing, interoperability and standardized formats, and dissimilar technological equipment adoption. Development of knowledge-based systems in the farming sector would require key components, supported by IoT, data acquisition systems, ubiquitous computing and networking, M2M communications, effective management of geospatial and temporal data, traceability systems along the supply chain, and ICT-supported cooperation among stakeholders. Recent designs and prototypes using cloud computing and the future Internet generic enablers for inclusion in FMIS have recently been proposed and lay the groundwork

for future applications. A modification, which is underway, from proprietary tools to Internet-based open systems supported by cloud hosting services will enable a more effective cooperation between actors of the supply chain. One of the limiting factors in the adoption of SCIS is a lack of interoperability, which would require implementation of virtual supply chains based on the virtualization of physical objects such as containers, products, and trucks. Recent and promising developments of the spatial decision-making deal with the interaction and the proactive involvement of the final users, implementing the so-called collaborative or participative Web-based GIS-MCDA systems. Computers science and IT evolvments affect the developments of RS in agriculture, leading to the need for new methods and solutions to the challenges of big data in a cloud computing environment.

Cross-References

- [Agriculture](#)
- [Cloud](#)
- [Data Processing](#)
- [Satellite Imagery/Remote Sensing](#)
- [Sensor Technologies](#)
- [Socio-spatial Analytics](#)
- [Spatial Data](#)

Further Reading

- Berckmans, D. (2004). Automatic on-line monitoring of animals by precision livestock farming. In *Proceedings of the ISAH conference on animal production in Europe: The Way Forward in a Changing World*. Saint-Malo, pp. 27–31.
- Li, M., & Chung, S. (2015). Special issue on precision agriculture. *Computers and Electronics in Agriculture*, 112, 1.
- Pierce, F. J., & Clay, D. (Eds.). (2007). *GIS applications in agriculture*. Boca Raton: CRC Press Taylor and Francis Group.
- Sørensen, C. G., Fountas, S., Nash, E., Pesonen, L., Bochtis, D., Pedersen, S. M., Basso, B., & Blackmore, S. B. (2010). Conceptual model of a future farm management information system. *Computers and Electronics in Agriculture*, 72(1), 37–47.

Agriculture

Madeleine Johnson, Hemayet Hossain,
Victor Sposito and Robert Faggian
Centre for Regional and Rural Futures, Deakin
University, Burwood, VIC, Australia

Synonyms

[AgInformatics](#); [Digital agriculture](#); [Smart agriculture](#)

Big Data and (Smart) Agriculture

Big data and digital technology are driving the latest transformation of agriculture – to what is becoming increasingly referred to as “smart agriculture” or sometimes “digital agriculture.” This term encompasses farming systems that employ digital sensors and information to support decision-making. Smart agriculture is an umbrella concept that includes precision agriculture (see ► [“AgInformatics”](#) – De Montis et al. 2017) – in many countries (e.g., Australia) precision agriculture commonly refers to cropping practices that use GPS guidance systems to assist with seed, fertilizer, and chemical applications. It therefore tends to be associated specifically with cropping farming systems and deals primarily with infield variability. Smart agriculture, however, refers to all farming systems and deals with decision-making informed by location, contextual data, and situational awareness. The sensors employed in smart agriculture can range from simple feedback systems, such as a thermostat that acts to regulate a machines temperature, to complex machine learning algorithms that inform pest and disease management strategies. The term big data, in an agricultural context, is related but distinct – it refers to computerized analytical systems that utilize large databases of information to identify statistical relationships that then inform decision support tools. This often includes big data from nonagricultural sources, such as weather or climate data or market data.

An example of how these concepts interact in practice: a large dataset may be established that contains the yield results of many varietal trials across a broad geographical area and over a long period of time (including detailed information pertaining to the location of each trial, such as soil type, climatic data, fertilizer, and chemical application rates, among others). This data could be analyzed to specifically determine the best variety for a particular geographic location and thus form the basis for a decision support system. These two steps constitute the data and the analytic components of “big data” in an agricultural context. The data could then inform other activities, such as the application (location and rate) of chemicals, fertilizers, and seed through digital-capable and GPS-guided farm machinery (precision agriculture).

Applications of Big Data in Smart Agriculture

Big data, and in particular big data analytics, are often described as disruptive technologies that are having a profound effect on economies. The amount of data being collected is increasing exponentially, and the cost of computing and digital sensors is decreasing exponentially. As such, the range of consumer goods (including farm machinery and equipment) that incorporates Internet or network connectivity as a standard feature is growing. The result is a rapidly expanding “Internet of things” (IoT) and large volumes of new data. For example, John Deere tractors are fitted with sensors that collect and transmit soil and crop data, which farmers can subscribe to access via proprietary software portals (Bronson and Knezevic 2016). The challenge in agriculture is reaching a point where available data and databases qualify as “big.” Yield measurements from a single paddock within one growing season are of little value because such limited data does not inform actionable decision taking. But, when the same data is collected across many paddocks and many seasons, it can be analyzed for trends that inform on-farm decision-making and thus

becomes much more valuable. This is true across the full agricultural value chain.

Nonetheless, smart agriculture, IoT, and big data are impacting on the full agricultural value chain. Here we list some examples according to farming system type (as outlined by AFI 2016):

1. Cropping systems: Variable rate application technology (precision agriculture), unmanned aerial vehicles or drones for crop assessment, remote sensing via satellite imagery.
2. Extensive livestock: Walkover weighing scales and auto-drafting equipment, livestock tracking systems, remote and proximal sensor systems for pasture management, virtual fencing.
3. Dairy: As for extensive livestock, plus individual animal ID systems and animal activity meters that both underpin integrated dairy and herd management systems.
4. Horticulture: Input monitoring and management systems (irrigation and fertigation), robotic harvesting systems, automated post-harvest systems (grading, packing, chilling).

Overall, while the technology is still relatively new, agriculture is already seeing substantial productivity gains from its use. Further transformative impacts will be felt when real-time information business process decisions and off-farm issues (e.g., postharvest track and trace of products), such as planning, problem-solving, risk management, and marketing are underpinned by big data.

Challenges and Implications

In an agricultural context, there are several challenges.

First, convincing farmers that the data (and its collection) are not merely a novelty but something that will drive significant productivity improvement in the future may be difficult. In many cases the hardware and infrastructure required to collect and use agricultural data are expensive (and prohibitively so in developing countries) or unavailable in rural areas (especially fast and

reliable Internet access), and the benefits may not be realized for many years. Similarly, technical literacy could be a barrier in some cases. These issues are, however, common to many on-farm practice change exercises that drive improvements in efficiency or productivity and can generally be overcome.

Second, farmers may perceive that there are privacy and security issues associated with making data about their farm available to unknown third parties (Wolfert et al. 2017). Large proprietary systems from the private sector are available to capture and store significant amounts of data that is then made available to farmers via subscription. But, linking big data systems to commercial benefit raises the possibility of biased recommendations. Similarly, farmers may be reluctant to provide detailed farm data to public or open-source decision support systems because they often do not trust government agencies. These systems also lack ongoing development and support for end users.

Finally, a sometimes-over-looked issue is that of data quality. In the race for quantity, it is easy to forget quality and the fact that not all digital sensors are created equal. Data is generated at varying resolutions, with varying levels of error and uncertainty, from machinery in various states of repair. The capacity of analytical techniques to keep pace with the amount of data, to filter out poor quality data, and to generate information that is suitable at a range of resolutions are all key issues for big analytics. For analyses to underpin accurate agricultural forecasting or predictive services that improve productivity, advancements in intelligent processing and analytics are required.

Ultimately, it is doubtful that farmer knowledge can ever be fully replaced by big data and analytic services. The full utility of big data for agriculture will be realized when the human components of food and fiber production chains are better integrated with the digital components to ensure that the outputs are relevant for planning (forecasting and predicting), communication, and management of (agri)business processes.

Cross-References

- [AgInformatics](#)
- [Data Processing](#)
- [Socio-spatial Analytics](#)
- [Spatial Data](#)

Further Reading

- Australian Farm Institute. (2016). *The implications of digital farming and big data for Australian agriculture*. Surry Hills: NSW Australian Farm Institute. ISBN 978-1-921808-38-8.
- Bronson, K., & Knezevic, I. (2016). Big data in food and agriculture. *Big Data & Society*, 3, 1–5. <https://doi.org/10.1177/2053951716648174>.
- De Montis, A., Modica, G., & Arcidiacono, C. (2017). AgInformatics. *Encyclopedia of Big Data*. https://doi.org/10.1007/978-3-319-32001-4_218-1.
- Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. (2017). Big data in smart farming – A review. *Agricultural Systems*, 153, 69–80. <https://doi.org/10.1016/j.agry.2017.01.023>.

AI

- [Artificial Intelligence](#)
-

Algorithm

Laurie A. Schintler¹ and Joshua Lee²

¹George Mason University, Fairfax, VA, USA

²Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Overview

We are now living in an “algorithm society.” Indeed, algorithms have become ubiquitous, running behind the scenes everywhere for various purposes, from recommending movies to optimizing autonomous vehicle routing to detecting fraudulent financial transactions. Nevertheless, algorithms are far from new. The idea of an

algorithm, referring generally to a set of rules to follow for solving a problem or achieving a goal, goes back thousands of years. However, the use of algorithms has exploded in recent years for a couple of interrelated reasons:

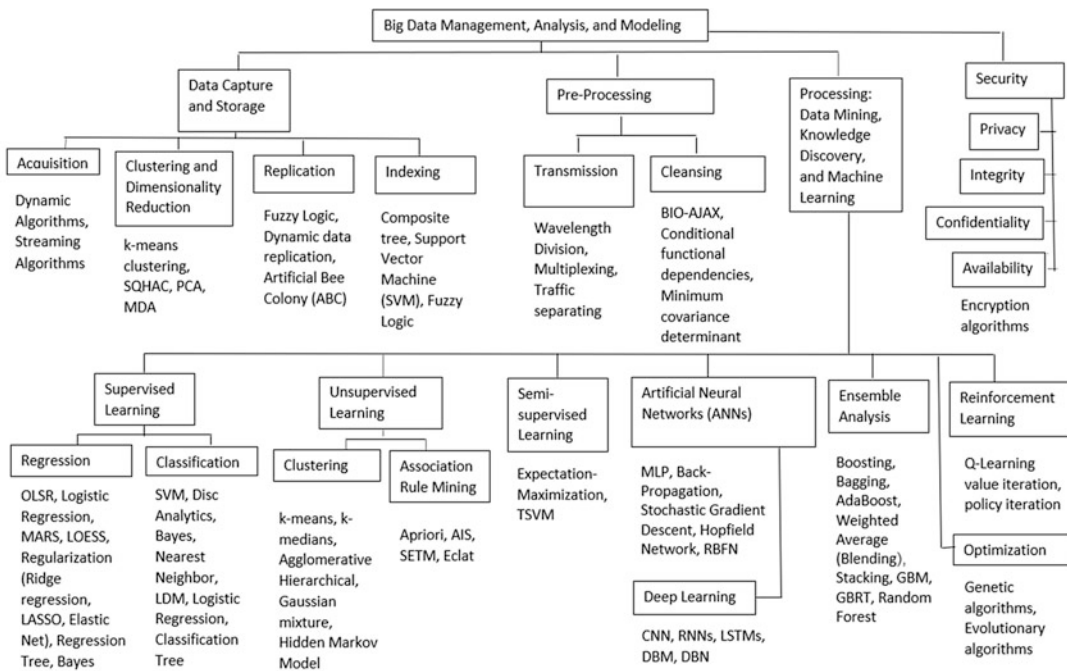
1. Advancements in computational and information processing technologies have made it easier to develop, codify, implement, and execute algorithms.
2. Open-source digital platforms and crowdsourcing projects enable algorithmic code to be shared and disseminated to a large audience.
3. The complexities and nuances of big data create unique computational and analytical challenges, which demand algorithms.

Algorithms used for big data management, analysis, modeling, and governance comprise a complex ecosystem, as illustrated in Fig. 1. Specifically, algorithms are used for capturing, indexing, and processing massive, fast-moving

data; extracting relevant and meaningful information and content from big data streams; detecting anomalies, predicting, classifying, and learning patterns of association; and protecting privacy and cybersecurity. Despite the benefits of algorithms in the big data era, their use and application in society come with various ethical, legal, and social downsides and dangers, which must be addressed and managed.

Machine Learning Algorithms

Machine learning leverages algorithms for obtaining insights (e.g., uncovering unknown patterns), creating models for prediction and classification, and controlling automated systems. In this regard, there are many different classes of algorithms. In supervised machine learning, algorithms are applied to a training data set containing attributes and outcomes (or labels) to develop a model that can predict or classify with a minimal level of model error. In contrast, unsupervised learning



Algorithm, Fig. 1 Ecosystem of algorithms for big data management, analysis, and modeling

algorithms are given a training set without any correct output (or labels) in advance. The algorithm's role is to figure out how to partition the data into different classes or groupings based on the similarity (or dissimilarity) of the attributes or observations. Association rule mining algorithms reveal patterns of association between features based on their co-occurrence. Semi-supervised algorithms are used in instances where not all observations have an output or label. Such algorithms exploit the available observations to create a partially trained model, which is then used to infer the output or labels for the incomplete observations. Finally, reinforcement learning algorithms, which are often used for controlling and maximizing automated agents' performance – e.g., autonomous vehicles, produce their own training data based on information collected from their interaction with the environment. Agents then adjust their behavior to maximize a reward or minimize risk.

Artificial Neural Networks (ANNs) are biologically inspired learning systems that simulate how the human brain processes information. Such models contain flexible weights along pathways connected to “neurons” and an activation function that shapes the nature of the output. In ANNs, algorithms are used to optimize the learning process – i.e., to minimize a cost function. Deep neural learning is an emerging paradigm, where the algorithms themselves adapt and learn the optimal parameter settings – i.e., they “learn to learn.” Deep learning contains many more layers and parameters than conventional ANNs. Each layer of nodes trains on features from the output of the prior layers. This idea, known as feature hierarchy, enables deep learning to effectively and efficiently model complex phenomena containing nonlinearities and multiple interacting features and dynamics.

Algorithms for Big Data Management

Conventional data management tools, techniques, and technologies were not designed for big data. Various kinds of algorithms are used to address

the unique demands of big data, particularly those relating to the volume, velocity, variety, veracity, and vulnerability of the data. Dynamic algorithms help to manage fast-moving big data. Specifically, such algorithms design data structures that reflect the evolving nature of a problem so that data queries and updates can be done quickly and efficiently without starting from scratch. As big data tends to be very large, it often exceeds our capacity to store, organize, and process it. Algorithms can be used to reduce the size and dimensionality of the data before it goes into storage and to optimize storage capacity itself. Big data tends to be fraught with errors, noise, incompleteness, bias, and redundancies, which can compromise the accuracy and efficiency of machine learning algorithms. Data cleansing algorithms identify imperfections and anomalies, transform the data accordingly, and validate the transformed data. Other algorithms are used for data integration, data aggregation, data transmission, data discretization, and other pre-processing tasks. A cross-cutting set of challenges relate to data security and, more specifically, the privacy, integrity, confidentiality, and accessibility of the data. Encryption algorithms, which encode data and information, are used to address such concerns.

Societal Implications of Algorithms

While algorithms are beneficial to big data management, modeling, and analysis, as highlighted, their use comes part and parcel with an array of downsides and dangers. One issue is algorithmic bias and discrimination. Indeed, algorithms have been shown to produce unfair outcomes and decisions, favoring (or disfavoring) certain groups or communities over others. The use of facial recognition algorithms for predicting criminality is a case and point. In particular, such systems are notoriously biased in terms of race, gender, and age. Algorithmic bias stems in part from the data used for training, testing, and validating machine learning models, especially if it is skewed or incomplete (e.g., due to sampling bias) or reflects

societal gaps and disparities in the first place. The algorithms themselves can also amplify and contribute to biases. Compounding matters are that algorithms are often opaque, particularly in deep learning models, which have complex architectures that cannot be easily uncovered, explained, or understood. Standards, policies, and ethical and legal frameworks are imperative for mitigating the negative implications of algorithms. Moreover, transparency is critical for ensuring that people understand the inner-workings of the algorithms that are used to make decisions that affect their lives and well-being. Considering new and advancing capabilities in Explainable Artificial Intelligence (XAI), algorithms themselves could soon play an active role in this regard, adding new dimensions and dynamics to the “algorithmic society.”

Cross-References

- [Algorithmic Complexity](#)
- [Artificial Intelligence](#)
- [Data Governance](#)
- [Deep Learning](#)
- [Machine Learning](#)

Further Reading

- Li, K. C., Jiang, H., Yang, L. T., & Cuzzocrea, A. (Eds.). (2015). *Big data: Algorithms, analytics, and applications*. Boca Raton: CRC Press.
- Mnich, M. (2018). Big data algorithms beyond machine learning. *KI – Künstliche Intelligenz*, 32(1), 9–17.
- Olhede, S. C., & Wolfe, P. J. (2018). The growing ubiquity of algorithms in society: Implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170364.
- Prabhu, C. S. R., Chivukula, A. S., Mogadala, A., Ghosh, R., & Livingston, L. J. (2019). Big data analytics. In *Big data analytics: Systems, algorithms, applications* (pp. 1–23). Singapore: Springer.
- Schuienburg, M., & Peeters, R. (Eds.). (2020). *The algorithmic society: Technology, power, and knowledge*. London: Routledge.
- Siddiq, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016).

A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications*, 71, 151–166.

Yu, P. K. (2020). The algorithmic divide and equality in the age of artificial intelligence. *Florida Law Review*, 72, 19–44.

Algorithmic Analysis

- [Algorithmic Complexity](#)

Algorithmic Complexity

Patrick Juola

Department of Mathematics and Computer Science, McNulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Synonyms

[Algorithmic analysis](#); [Big O notation](#)

Introduction

Algorithmic complexity theory is the theoretical analysis of the amount of resources consumed by a process in executing a particular algorithm or solving a particular problem. As such, it is a measure of the inherent difficulty of various problems and also of the efficiency of proposed solutions. The resources measured can be almost anything, such as the amount of computer memory required, the number of gates required to embed the solution in hardware, and the number of parallel processors required, but it most often refers to the amount of time required for a computer program to successfully execute and, in particular, to differences in the amount of resources that cannot be solved simply by using better equipment.

An Example

Consider the problem of determining whether each element in an N -element array is unique or, in other words, whether or not the array contains any pairs. A naïve but simple solution would be to compare every element with every other element; if no two elements are equal, every element is unique. The following pseudocode illustrates this algorithm:

Algorithm 1:

```

for every element  $a[i]$  in the array
  for every element  $a[j]$  in the array
    if  $i \neq j$  and  $a[i] = a[j]$  report false and
quit (Statement A1)
  if all element-pairs have been compared,
report true and quit

```

Because there are N^2 element-pairs to compare, Statement A1 will be executed up to N^2 times. The program as a whole will thus require at least N^2 statement execution times to complete.

A slightly more efficient algorithm designer would notice that if element $a[x]$ has been compared to element $a[y]$, there is no need to compare element $a[y]$ to element $a[x]$ later. One can therefore restrict comparisons between element $a[x]$ and elements later in the array, as in the following pseudocode:

Algorithm 2:

```

for every element  $a[i]$  in the array
  for every element  $a[j]$  ( $j > i$ ) in the array
    if  $a[i] = a[j]$  report false and quit
(Statement A2)
  if all element-pairs have been compared,
report true and quit

```

In this case, the first element will be compared against $N-1$ other elements, the second against $N-2$, and so forth. Statement A2 will thus be executed $(1 + 2 + 3 + 4 + \dots + (N-1))$ times, for a total of $N(N-1)/2$ times. Since $N^2 > N(N-1)/2$, Algorithm 2 could be considered marginally more efficient. However, note that this comparison assumes that Algorithms 1 and 2 are running on

comparable computers. If Algorithm 1 were run on a computer 10 times as fast, then it would complete in (effectively) time equal to $N^2/10$, faster than Algorithm 2.

By contrast, Algorithm 3 is inherently more efficient than either of the other algorithms, sufficiently faster to beat any amount of money thrown at the issue:

Algorithm 3:

```

sort the array such that  $a[i] \geq a[i + 1]$  for
every element  $i$  (Statement A3)
  for every element  $a[i]$  in the (sorted) array
    if  $a[i] = a[i + 1]$  report false and quit
(Statement A4)
  if all element-pairs have been compared,
report true and quit

```

The act of sorting will bring all like-valued elements together; if there are pairs in the original data, they will be in adjacent elements after sorting, and a single loop looking for adjacent elements with the same value will find any pairs (if they exist) in N passes or fewer through the loops. The total time to execute Algorithm 3 is thus roughly equal to N (the number of times statement A4 is executed) plus the amount of time it takes to sort an array of N elements.

Sorting is a well-studied problem; many different algorithms have been proposed, and it is accepted that it takes approximately N times $\log_2(N)$ steps to sort such an array. The total time of Algorithm 3 is thus $N + N \log_2(N)$, which is less than $2(N \log_2(N))$, which in turn is less than N^2 for large values of N . Algorithm 3, therefore, is more efficient than Algorithm 1 or 2, and the efficiency gap gets larger as N (the amount of data) gets bigger.

Mathematical Expression

Complexity is usually expressed in terms of complexity classes using the so-called algorithmic notation (also known as “big O” or “big Oh” notation.) In general, algorithmic notation describes the limit behavior of a function in terms of equivalence classes. For polynomial

functions, such as $aN^3 + bN^2 + cN^1 + d$, the value of the function is dominated (for large N) by N^3 . If $N \gg a$, then the exact value of a does not matter very much, and even less do the values of b , d , and d . Similarly, for large N , any (constant) multiplier of N^2 is larger than any constant times $N \log_2 N$, which in turn is larger than any constant multiplier of N .

More formally, for any two functions $f(N)$ and $g(N)$,

$$f(N) = O(g(N)) \quad (1)$$

if and only if there are positive constants K and n^0 such that

$$|f(N)| \leq K|g(N)| \text{ for all } N > n_0 \quad (2)$$

In less formal terms, as N gets larger, a multiple of the function $g()$ eventually gets above $f()$ and stays there indefinitely. Thus, even if you speeded up the algorithm represented by $f()$ by any constant multiplier (e.g., by running the program on a computer K times as fast), $g()$ would still be more efficient for large problems.

Because of the asymmetry of this definition, the $O()$ notation specifically establishes an upper bound (worst case) on algorithm efficiency. There are other, related notations (“big omega” and “big theta”) that denote upper bounds and exact (upper and lower) bounds, respectively.

In practice, this definition is rarely used; instead people tend to use a few rules of thumb to simplify calculations. For example, if $f(N)$ is the sum of several terms, only the largest term (the one with the largest power of N) is of interest. If f is the product of several factors, only factors that depend on x are of interest. Thus if $f()$ were the function

$$f(N) = 21N^3 + 3N^2 + 17N - 4 \quad (3)$$

the first rule tells us that only the first term ($21N^3$) matters, and the second rule tells us that the constant 21 does not matter. Hence

$$f(N) = O(N^3) \quad (4)$$

as indeed would any cubic polynomial function.

An even simpler rule of thumb is that the deepest number of nested loops in a computer program or algorithm controls the complexity of the overall program. A program that loops over all the data will need to examine each point and hence is at least $O(N)$. A program that contains two nested loops (such as Algorithms 1 and 2) will be $O(N^2)$ and so forth.

Some Examples

As discussed above, the most naïve sorting algorithms are $O(N^2)$ as they involve comparing each time to most if not all other items in the array. Fast sorting algorithms such as *mergesort* and *heapsort* are $O(N \log_2(N))$. Searching for an item in an unsorted list is $O(N)$ because every element must potentially be examined. Searching for an item in a sorted list is $O(N \log_2(N))$ because binary search can be used to eliminate large sections of the list.

Problems that can be solved in constant time (such as determining if a number is positive or negative) are said to be $O(1)$.

A particularly important class of algorithms are those for which the fastest known algorithm is exponential ($O(c^N)$) or worse. For example, the so-called travelling salesman problem involves finding the shortest closed path through a given set of points. These problems are generally considered to be very hard to solve as the best-known algorithm is still very complex and time-consuming.

Further Reading

- Aho, A. V., & Ullman, J. D. (1983). *Data structures and algorithms*. Pearson.
- Knuth, D. (1976, Apr-June). Big omicron and big omega and big theta. *SIGACT News*.
- Knuth, D. E. (1998). *Sorting and searching, 2nd edn. The art of computer programming, vol. 3* (p. 780). Pearson Education.
- Sedgewick, R., & Wayne, K. (2011). *Algorithms*. Addison-Wesley Professional.

American Bar Association

Jennifer J. Summary-Smith

Florida SouthWestern State College, Fort Myers,
FL, USA

Culver-Stockton College, Canton, MO, USA

The American Bar Association (ABA) is one of the world's largest voluntary associations of lawyers, law students, legal, and law professions in the United States. Its national headquarters is located in Chicago, Illinois, with a large branch office in Washington D.C. According to the ABA's website, it has nearly 400,000 members and more than 3,500 entities. The American Bar Association was established when 75 lawyers from 20 states, and the District of Columbia, came together on August 21, 1878, in Saratogo Springs, New York. Since its founding in 1878, the ABA has played an important role in the development of the law profession in the United States. The ABA website states that "the ABA is committed to supporting the legal profession with practical resources for legal professionals while improving the administration of justice, accrediting law schools, establishing model ethical codes, and more." The ABA is also committed to serving its members, refining the legal profession, eradicating bias and promoting diversity, and evolving the rule of law in the entire United States and around the globe. Thus, becoming a member of ABA has several benefits in terms of access to exclusive data.

Membership Benefits Involving Big Data

A benefit of becoming a member of the ABA is that it allows access to free career services for job seekers and employers. As it states on the ABA's website, job seekers can search and apply for more than 450 legal jobs across the nation. The ABA's website provides the opportunity to upload one's resume, receive email alerts, and access monthly webinars by experts who provide career advice. Employers have access to more than 5,400 resumes, email alerts, and reach more than 16,500

visitors monthly. The career services provide members the opportunity to network with potential employers, granting access to valuable data and personal information.

Other benefits for members include access to the ABA's 22 sections, 6 divisions, and 6 forums. Members can participate in a community where they can interact with professionals in a variety of practice specialties. Each of the groups provides members the opportunity to facilitate in-depth examinations of trends, issues, and regulations in specific areas of law and other special interests. Members can also enrich their careers with ABA's committees and task forces, which provide access to specialty groups and internal ABA departments; the groups range from anti-trust and family law, to the law student division and judicial division. The ABA also advocates for inclusion and diversity initiatives committed to eliminating bias and promoting diversity. The ABA publishes annual reports on the following issues: persons with disabilities participation rates, racial and ethnic diversity, women in leadership positions, and lesbian, gay, bisexual, and transgender participation. Through the use of the ABA's data, members are able to learn and understand valuable information, regarding the every changing landscape of law and society. Members use the data to help guide their own practices, influencing decision-making and public policy. Although technology can positively affect ABA members' careers by providing vital information, there are concerns in the legal profession in regard to its influence on social interaction within the work environment.

Benefits and Concerns of Technology in the Workplace

In a recent study by Glen Vogel, he analyzed issues associated with the generational gap in using technology and social media by legal professionals. According to the article, Internet social media (ISM) is a concern within the profession, blurring the lines between professional and personal tasks. ISM can also foster technology overload, resulting a need for reevaluating workplace

etiquette and rules of professional conduct. Vogel posits that over the past decade legal professionals have been using ISM for more than connecting with people. Users are participating on a global front, engaging in ISM to influence society. With a surge in the amount of users in the legal workplace, there are growing concerns with confidentiality and the traditional work environment. As younger generations enter the workforce, the gap between older generations widens. Vogel adds it is important for every generation to be willing to accept new technologies because they can provide to be useful tools within the workplace.

The American Bar Association is a proprietor of big data, influencing the legal profession in the United States and around the world. The ABA continues to expand its information technology services, recently partnering with ADAR IT. Marie Lazzara writes that ADAR is the provider of the private cloud, supporting law firms with benefits such as remote desktop access and disaster recovery. As more organizations, such as the ABA, make strides to bridge this gap, one thing is certain, the big data phenomenon has an influence on the legal profession.

Cross-References

- Cloud Services
- Data Brokers
- Ethical and Legal Issues
- LexisNexis

Further Reading

- American Bar Association, <http://www.americanbar.org/aba.html>. Accessed July 2014.
- Lazzara, M. ADAR IT Named Premium Solutions Provider by American Bar Association. <http://www.prweb.com/releases/2014/ADAR/prweb12053119.htm>. Accessed July 2014.
- Vogel, G. (2013). A Review of the International Bar Association, LexisNexis Technology Studies, and the American Bar Association's Commission on Ethics 20/20: the Legal Profession's Response to the Issues Associated With the Generational Gap in Using Technology and Internet Social Media. *The Journal of the Legal Profession*, 38, 95 p.

American Civil Liberties Union

Doug Tewksbury
Communication Studies Department,
Niagara University, Niagara, NY, USA

The American Civil Liberties Union (ACLU) is an American legal advocacy organization that defends US Constitutional rights through civil litigation, lobbying efforts, educational campaigns, and community organization. While not its sole purpose, the organization has historically focused much of its attention on issues surrounding the freedom of expression, and as expression has become increasingly mediated through online channels, the ACLU has fought numerous battles to protect individuals' First and Fourth Amendment rights of free online expression, unsurveilled by government or corporate authorities.

Founded in 1920, the ACLU has been at the forefront of a number of precedent-setting cases in the US court system. It is perhaps most well-known for its defense of First Amendment rights, particularly in its willingness to take on unpopular or controversial cases, but has also regularly fought for equal access and protection from discrimination (particularly for groups of people who have traditionally been denied these rights under the law), Second Amendment protection for the right to bear arms, and due process under the law, amongst others. The ACLU has provided legal representation or *amicus curiae* briefs for a number of notable precedent-setting legal cases, including *Tennessee v. Scopes* (1921), *Gitlow v. New York* (1925), *Korematsu v. United States* (1944), *Brown v. Board of Education* (1954), *Miranda v. Arizona* (1966), *Roe v. Wade* (1973), and dozens of others. Its stated mission is "to defend and preserve the individual rights and liberties guaranteed to every person in this country by the Constitution and laws of the United States."

The balance between civil liberties and national security is an always-contentious relationship, and the ACLU has come down strongly

on the side of privacy for citizens. The passage of the controversial USA PATRIOT Act in 2001 and its subsequent renewals led to sweeping governmental powers of warrantless surveillance, data collection, wiretapping, and data mining, many of which continue to today. Proponents of the bill defended its necessity in the name of national security in the digital age; opponents argued that it would fundamentally violate the civil rights of American citizens and create a surveillance state. The ACLU would be among the leading organizations challenging a number of practices resulting from the passage of the bill.

The cases during this era are numerous, but several are particularly noteworthy in their relationship to governmental and corporate data collection and surveillance. In 2004, the ACLU represented Calyx Internet Access, a New York internet service provider, in *Doe v. Ashcroft*. The FBI had ordered the ISP to hand over user data through issuing a National Security Letter, a de facto warrantless subpoena, along with issuing a gag order on discussing the existence of the inquiry, a common provision of this type of letter. In *ACLU v. National Security Agency* (NSA) (2006), the organization unsuccessfully led a lawsuit against the federal government arguing that its practice of warrantless wiretapping was a violation of Fourth Amendment protections. Similarly lawsuits were filed against AT&T, Verizon, and a number of other telecommunication corporations during this era. The ACLU would represent the plaintiffs in *Clapper v. Amnesty International* (2013), an unsuccessful attempt to challenge the Foreign Intelligence Surveillance Act's provision that allows for the NSA's warrantless surveillance and mass data collection and analysis of individuals' electronic communications. It has strongly supported the whistleblower revelations of Edward Snowden in his 2013 leak of classified NSA documents detailing the extent of the organization's electronic surveillance of the communications of over a billion people worldwide, including millions of domestic American citizens.

In terms of its advocacy campaigns, the organization has supported *Digital 4th*, a Fourth Amendment activist group, advocating for a non-partisan focus on new legislative action to update the now-outdated Electronic Communications Privacy Act (ECPA), a 30-year-old bill that still governs much of online privacy law. Similarly, the ACLU has strongly supported Net Neutrality, the equal distribution of high-speed broadband traffic. The *Free Future* campaign has made the case for governmental uses of technology in accountable, transparent, and constitutionally sound ways on such issues as body-worn cameras for police, digital surveillance and data mining, hacking and data breaches, and traffic cameras, amongst other technological issues, as well as through the *Demand Your DotRights* campaign. In 2003, the CAN-SPAM act was on its way through Congress, and the ACLU took the unpopular position that the act unjustly restricted the freedom of speech online and would serve as a chilling effect on speech, as it has continued to argue in several other anti-spam legislative bills.

The ACLU has built its name on defending civil rights, and the rise of information-based culture has resulted in a greatly expanded practice and scope of the organization's focus. However, with cases such as the 2013–2014 revelations that came from the Snowden affair on NSA surveillance, it is clear that the ongoing tension between the rise of new information technologies, the government's desire for surveillance in the name of national security, and the public's right to Constitutional protection under the Fourth Amendment is far from resolved.

Further Reading

- American Civil Liberties Union. (2014). Key Issues/About Us. Available at <https://www.aclu.org/key-issues>.
- Herman, S. N. (2011). *Taking liberties: The war on terror and the erosion of American democracy*. New York: Oxford University Press.
- Klein, W., & Baldwin, R. N. (2006). *Liberties lost: The endangered legacy of the ACLU*. Santa Barbara: Greenwood Publishing Group.
- Walker, S. (1999). *In defense of American liberties: A history of the ACLU*. Carbondale, IL: SIU Press.

American Library Association

David Brown

Southern New Hampshire University, University of
Central Florida College of Medicine, Huntington
Beach, CA, USA

University of Wyoming, Laramie, WY, USA

The American Library Association (ALA) is a voluntary organization that represents libraries and librarians around the world. Worldwide, the ALA is the largest and oldest professional organization for libraries, librarians, information science centers, and information scientists. The association was founded in 1876 in Philadelphia, Pennsylvania. Since its inception, the ALA has provided leadership for the development, promotion, and improvement of libraries, information access, and information science. The ALA is primarily concerned with learning enhancement and information access for all people. The organization strives to advance the profession through its initiatives and divisions within the organization. The primary action areas for the ALA are advocacy, education, lifelong learning, intellectual freedom, organizational excellence, diversity, equitable access to information and services, expansion of all forms of literacy, and library transformation to maintain relevance in a dynamic and increasing global digitalized environment. While ALA is composed of several different divisions, there is no single division devoted exclusively to big data. Rather, a number of different divisions are working to develop and implement policies and procedures that will enhance the quality of, the security of, the access to, and the utility of big data.

ALA Divisions Working with Big Data

At this time, the Association of College & Research Libraries (ACRL) is a primary division of the ALA that is concerned with big data issues. The ACRL has published a number of papers, guides,

and articles related to the use of, promise of, and the risks associated with big data. Several other ALA divisions are also involved with big data. The Association for Library Collections & Technical Service (ALCTS) division discusses issues related to the management, organization, and cataloging of big data and its sources. The Library Information Technology Association (LITA) is an ALA division that is involved with the technological and user services activities that advance the collection, access, and use of big data and big data sources.

Big Data Activities of the Association of College & Research Libraries (ACRL)

The Association of College & Research Libraries (ACRL) is actively involved with the opportunities and challenges presented by big data. As science and technology advance, our world becomes more and more connected and linked. These links in and of themselves may be considered big data, and much of the information that they transmit is big data. Within the ACRL, big data is conceptualized in terms of the three Vs: its volume, its velocity, and its variety. Volume refers to the tremendously large size of the big data. However, ACRL stresses that the size of the data set is a function of the particular problem one is investigating and size is only one attribute of big data. Velocity refers to the speed at which data is generated, needed, and used. As new information is generated exponentially, the need to catalogue, organize, and develop user-friendly means of accessing these big data increases multiple exponentially. The utility of big data is a function of the speed at which it can be accessed and used. For maximum utility, big data needs to be accurately catalogued, interrelated, and integrated with other big data sets. Variety refers to the many different types of data that are typically components of and are integrated into big data. Traditionally, data sets consist of a relatively small number of different types of data, like word-processed documents, graphs, and pictures. Big data on the other hand is typically concerned with many additional types

of information such as emails, audio and videotapes, sketches, artifacts, data sets, and many other kinds of quantitative and qualitative data. In addition, big data information is usually presented in many different languages, dialects, and tones. A key point that ACRL stresses is that as disciplines advance, the need for and the value of big data will increase. However, this advancement can be facilitated or inhibited by the degree to which the big data can be accessed and used. Within this context, librarians who are also information scientists are and will continue to be invaluable resources that can assist with the collection, storage, retrieval, and utilization of big data. Specifically, ACRL anticipates needs for specialists in the areas of big data management, big data security, big data cataloguing, big data storage, big data updating, and big data accessing.

Conclusion

The American Library Association and its member libraries, librarians, and information scientists are involved in shaping the future of big data. As disciplines and professions continue to advance with big data, librarians and information scientists' skills need to advance to enable them to provide valuable resources for strategists, decision-makers, policy-makers, researchers, marketers, and many other big data users. The ability to effectively use big data will be a key to success as the world economy and its data sources expand. In this rapidly evolving environment, the work of the ALA will be highly valuable and an important human resource for business, industry, government, academic and research planners, decision-makers, and program evaluators who want and need to use big data.

Cross-References

- [Automated Modeling/Decision Making](#)
- [Big Data Concept](#)
- [Big Data Quality](#)
- [Data Preservation](#)
- [Data Processing](#)
- [Data Storage](#)

Further Reading

- American Library Association. About ALA. <http://www.ala.org/aboutala/>. Accessed 10 Aug 2014.
- American Library Association. Association for Library Collections and Technical Services. <http://www.ala.org/alcts/>. Accessed 10 Aug 2014.
- American Library Association. Library Information Technology Association (LITA). <http://www.ala.org/lita/>. Accessed 10 Aug 2014.
- Bieraugel, Mark. Keeping up with... big data. American Library Association. http://www.ala.org/acrl/publications/keeping_up_with/big_data. Accessed 10 Aug 2014.
- Carr, P. L. (2014). Reimagining the library as a technology: An analysis of Ranganathan's five laws of library science within the social construction of technology framework. *The Library Quarterly*, 84(2), 152–164.
- Federer, L. (2013). The librarian as research informationist: A case study. *Journal of the Medical Library Association*, 101(4), 298–302.
- Finnemann, N. O. (2014). Research libraries and the Internet: On the transformative dynamic between institutions and digital media. *Journal of Documentation*, 70(2), 202–220.
- Gordon-Murnane, L. (2012). Big data: A big opportunity for Librarians. *Online*, 36(5), 30–34.

Animals

Marcienne Martin

Laboratoire ORACLE [Observatoire Réunionnais des Arts, des Civilisations et des Littératures dans leur Environnement] Université de la Réunion
Saint-Denis France, Montpellier, France

If in the digital world, an array of data in exponential growth is compiled, as expressed by Microsoft in the following: “data volume is expanding tenfold every five years. Much of this new data is driven by devices from the more than 1.2 billion people who are connected to the Internet worldwide, with an average of 4.3 connected devices per person (Microsoft Modern Data Warehouse white paper.pdf (2016, p. 6) – <https://www.microsoft.com/fr-fr/sql-server/big-data-data-warehousing>),” their redistribution varies according to the topic concerned. Thus, the animal world can be broken down according to a descriptive and analytical

mode (biology, for example) but also through the emotional field of the human being.

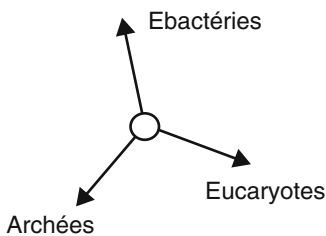
The living world is based on the synthesis of complex molecular developments which have evolved towards an autocatalytic system, reproductive and evolutionary (Calvin). Darwin was the precursor of many studies on the origin and evolution of species. In this regard, Philippe et al. indicate (1995) that the present species contain in their genome sequences inherited from a common progenitor. Eukaryotes form a set of lines, in which – with animals, plants, and fungi – all the great biological groups are found. For the majority of us, these groups appear to constitute the majority of the diversity of the living world and, moreover, contain our own species. This arborescent type structure is shown in the diagram below (Lecointre and Le Guyader 2001) (Fig. 1).

Communication, in whatever form, is the substructure that allows various species of the world of the living to continue to exist in space and in time. The transmission of information is also part such as spotting means. At the same time, predator and prey, living developed its way of life through the search for food and through its own protection as well as that of its species. This functioning mode corresponds to the level 1 of Maslow's pyramid of needs, which is basic needs, like food or shelter. With the emergence of language in the hominid, primate, member of the simian group, communication started to use other tools. Indeed, the particularity of the human being is his or her thoughts, more precisely his or her (=their) consciousness of their existence. This was affirmed by Descartes (2000) in his famous formula: *Cogito ergo sum*. The thought is

associated with the language whose intentionality serves the adaptation of the *Homo sapiens* to their environment through the creation and transmission of informative messages given to their congeners.

In addition, both cognitive and language structures are subdivided into various layers, such as the representation of objects in the world and their symbolization. The relation between humans/animals in their function of predator/prey is the basis of a reconstruction of the animal by the human being as part of a symbolic approach. Lévi-Strauss, anthropologist, demonstrated that the concept of totem was born from a human being's identification with certain animal characteristics as among the Chippawa tribe, North American Indians, where people of the "fish clan" had little hair, those of the "bear clan" were distinguished by long, black hair and an angry and combative temperament, or those of the "clan of the crane" by a screaming voice (1962, p. 142). In contrast, we find anthropomorphized animals in some fairy tales, such as "Little Red Riding Hood" by Grimm where the wolf plays the role of a carnivorous grandmother or in fables, like those of La Fontaine.

The imaginary for humans has contributed to the reconstruction of the animal as part of the Greek mythology, such as the Centaurs, hybrid beings, half human, half *Equidae*, or Medusa, one of the three Gorgons, whose hair was made of snakes. Some divine entities wear accessories belonging to the animal world as the devil with horns worn by *Bovidae* or the angels with their wings referring to the species of birds. Superstition gives some animals protective or destructive powers, such as a black cat which was associated with witchcraft in the Middle Ages, or at that times, when human beings found a swarm of bees attached to a tree in their garden this phenomenon was considered a bad sign they had to give a silver coin to these insects as a New Year's gift (Lacarrière 1987). The sacralization of the animal is also a special relationship of the human being with animals, like the bull-headed god Apis, or the sacred cat in ancient Egypt. Caricatures have been also inspired by animals to highlight particular character traits at such a known



Animals, Fig. 1 Diagram of the living world

public personality. The projection of the human being and animals between human entry in a register other than his own species or that of the animal in the human species may be born out of telescoping of predator and prey roles played by all living and questioning the Human being. Modern technologies are at the origin of a new animal mythology with well-known animated films, such as those of Walt Disney and its various characters, such as Mickey Mouse or Donald Duck.

The representation of an object of the world evolves according to various factors, including the progress of science. Various studies have tried to understand the mode of thinking in the animal in comparison with that of the human being. Dortier (1998) specifies as well as everywhere in the living world that animals exhibit more or less elaborated cognitive abilities. Furthermore, primatology, which is the science dedicated to the study of the species of primates, shows in the context of the phylogenetic filiations of the pygmy chimpanzees of Zaire and African chimpanzees that we share 98% of their genetic program (Diamond 1992, p. 10). This new approach to the human being in relation to animals where it mentions his belonging to the animal world may have changed the perception regarding the animal world. The protection of the animal, which is considered a sensitive being, has become wide-spread in the societies of the twenty-first century.

In its relation with the human being, the term animal includes two categories: the wild animal and the domestic animal. The latter lives on the personal territory of the human being and also enters their emotional field. In a search made with the help of the Google (<https://www.google.fr/search?q=hashtag&oq=Hastag&aqs=chrome.69i57j0l5.3573j0j7&sourceid=chrome&ie=UTF-8#q=twitter+animaux>) search engine, the number of sites which express themselves with Twitter (<https://twitter.com/?lang=fr>) – a service which is used to relay short information from user to user approximate the figure of 32,400,000 results. It is worth noting that the term “twitter”

refers to the different songs emitted by birds (class of the Aves). Applications, such as Hashtag (<https://fr.wikipedia.org/wiki/Hashtag>), which is a “meaningful continuation sequence of written characters without a space, beginning with the # sign (sharp) (<http://www.programme-tv.net/news/buzz/44259-twitter-c-est-quoi-un-hashtag/>),” YouTube (<https://www.youtube.com/?gl=FR&hl=fr>), which offers every user to create videos and put them online, allowing any Internet user to share their different experiences, whatever their nature, in their relationship with animals, or, again, Instagram (<https://www.instagram.com/?hl=fr>), which opens up the sharing of photos and videos between friends. An example that made the buzz on Instagram is that of Koyuki, the grumpy cat (<https://fr.pinterest.com/pin/553802085412023724/>).

Further Reading

- Calvin, M. (1975). *L'origine de la vie. La recherche en biologie moléculaire* (pp. 201–222). Paris: Editions du Seuil.
- Darwin, C. (1973). *L'origine des espèces*. Verviers: Marabout Université.
- Descartes, R. (2000). *Discours de la méthode*. Paris: Flammarion.
- Diamond, J. (1992). *Le troisième singe – Essai sur l'évolution et l'avenir de l'animal humain*. Paris: Gallimard.
- Dortier, J. F. (1998). *Du calamar à Einstein... L'évolution de l'intelligence. Le cerveau et la pensée – La révolution des sciences cognitives* (pp. 303–309). Paris: Éditions Sciences humaines.
- Lacarrière, J. (1987). *Les évangiles des quenouilles*. Paris: Imago.
- Lecointre, G., & Le Guyader, H. (2001). *Classification phylogénétique du vivant*. Paris: Belin.
- Lévi-Strauss, C. (1962). *La pensée sauvage*. Paris: Librairie Plon.
- Maslow, A. (2008). *Devenir le meilleur de soi-même – Besoins fondamentaux, motivation et personnalité*. Paris: Eyrolles.
- Philippe, H., Germot, A., Le Guyader, H., & Adoutte, A. (1995). Que savons-nous de l'histoire évolutive des eucaryotes ? 1. L'arbre universel du vivant et les difficultés de la reconstruction phylogénétique. *Med Sci*, 11, 8 (I–XIII), 1–2. http://www.ipubli.inserm.fr/bitstream/handle/10608/2438/MS_1995_8_I.pdf.

Anomaly Detection

Feras A. Batarseh
College of Science, George Mason University,
Fairfax, VA, USA

Synonyms

Defect detection; Error tracing; Testing and evaluation; Verification

Definition

Anomaly Detection is the process of uncovering anomalies, errors, bugs, and defects in software to eradicate them and increase the overall quality of a system. Finding anomalies in big data analytics is especially important. Big data is “*unstructured*” by definition, hence, the process of *structuring* it is continually presented with anomaly detection activities.

Introduction

Data engineering is a challenging process. Different stages of the process affect the outcome in a variety of ways. Manpower, system design, data formatting, variety of data sources, size of the software, and project budget are among the variables that could alter the outcome of an engineering project. Nevertheless, software and data anomalies pose one of the most challenging obstacles in the success of any project. Anomalies have postponed space shuttle launches, caused problems for airplanes, and disrupted credit card and financial systems. Anomaly detection is commonly referred to as a science as well as an art. It is clearly an inexact process, as no two testing teams will produce the same exact testing design or plan (Batarseh 2012).

Anomaly Examples

The cost of failed software can be high indeed. For example, in 1996, a test flight of a European launch system, *Ariane 5 # 501*, failed as a result of an anomaly. Upon launch, the rocket veered off its path and was destroyed by its self-destruction system to avoid further damage. This loss was later analyzed and linked to a simple floating number anomaly. Another famous example is regarding a wholesale pharmaceutical distribution company in Texas (called: Fox Meyer Drugs). The company developed a resources planning system that failed right after implementation, because the system was not tested thoroughly. When Fox Meyer deployed the new system, most anomalies floated to the surface, and caused lots of users’ frustration. That put the organization into bankruptcy in 1996. Moreover, three people died in 1986 when a radiation therapy system called Therac erroneously subjected patients to lethal overdoses of radiation. More recently however, in 2005, Toyota recalled 160,000 Prius automobiles from the market because of a software anomaly in the car’s software. The mentioned examples are just some of the many projects gone wrong (Batarseh and Gonzalez 2015); therefore, anomaly detection is a critical and difficult issue to address.

Anomaly Detection Types

Although anomalies can be prevented, it is not an easy task to build fault-free software. Anomalies are difficult to trace, locate, and fix; they can occur due to multiple reasons, examples include: due to a programming mistake, miscommunication among the coders, a misunderstanding between the customer and the developer, a mistake in the data, error in the requirements document, a politically biased managerial decision, a change in the domain market standards, and multiple other reasons. In most cases, however, anomalies fall under one of the following categories (Batarseh 2012):

1. **Redundancy** – Having the same data in two or more places.
2. **Ambivalence** – Mixed data or unclear representation of knowledge.
3. **Circularity** – Closed loops in software; a function or a system leading to itself as a solution.
4. **Deficiency** – Inefficient representation of requirements.
5. **Incompleteness** – Lack of representation of the data or the user requirements.
6. **Inconsistency** – Any untrue representation of the expert's knowledge.

Different anomaly detection approaches that have been widely used in many disciplines are presented and described in the Table 1.

However, based on the recent study by National Institute of Standards and Technology (NIST), the data anomaly itself is not the quandary, it is actually the ability to identify the **location of the anomaly**. That is listed as the most time-consuming activity of testing. In their study, NIST researchers compiled a vast number of software and data projects and reached the following conclusion: “If the location of bugs can be made more precise, both the calendar time and resource requirements of testing can be reduced. Modern data and software products typically contain millions of lines of code. Precisely locating the source of bugs in that code can be very resource consuming.” Based on that, it can be concluded that anomaly detection is an important area of research that is worth exploring (NIST 2002; Batarseh and Gonzalez 2015).

Conclusion

Similar to most engineering domains, software and data require extensive testing and evaluation. The main goal of testing is to eliminate anomalies, in a process referred to as anomaly detection.

It is not possible to perform data analysis if the data has anomalies. Data scientists usually perform steps such as data cleaning, aggregation, filtering, and many others. All these activities require anomaly detection to be able to verify

Anomaly Detection, Table 1 Anomaly detection approaches

Anomaly detection approach	Short description
Detection through analysis of heuristics	Logical validation with uncertainty, a field of artificial intelligence
Detection through simulation	Result-oriented validation through building simulations of the system
Face/field validation and verification	Preliminary approach (used with other types of detection). This is a usage-oriented approach
Predictive detection	A software engineering method, part of testing
Subsystem testing	A software engineering method, part of testing
Verification through case testing	Result-oriented validation, achieved by running tests and observing the results
Verification through graphical representations	Visual validation and error detection
Decision trees and directed graphs	Visual validation – observing the trees, and the structure of the system
Simultaneous confidence intervals	Statistical/quantitative verification
Paired T-tests	Statistical/quantitative verification
Consistency measures	Statistical/quantitative verification
Turing testing	Result-oriented validation, one of the commonplace artificial intelligence methods
Sensitivity analysis	Result-oriented data analysis
Data collection and outlier detection	Usage-oriented validation through statistical methods and data mining
Visual interaction verification	Visual validation through user interfaces

the data and provide valid outcomes. Additionally, detection leads to a better overall quality of a data system, therefore, it is a necessary and an unavoidable process. Anomalies occur for many reasons, and in many parts of the system, many practices lead to anomalies (listed in this entity), locating them, however, is an interesting engineering problem.

Cross-References

► Data Mining

Further Reading

- Batarseh, F. (2012). *Incremental lifecycle validation of knowledge-based systems through CommonKADS*. Ph.D. Dissertation Registered at the University of Central Florida and the Library of Congress.
- Batarseh, F., & Gonzalez, A. (2015). *Predicting failures in contextual software development through data analytics*. Proceedings of Springer's Software Quality Journal.
- Planning Report for NIST. (2002). *The economic impacts of inadequate infrastructure for software testing*. A report published by the US Department of Commerce.

Anonymity

Pilar Carrera

Universidad Carlos III de Madrid, Madrid, Spain

What matter who's speaking (Beckett)

"Anonymity" refers to the quality or state of being anonymous, from Greek *anonymos* and Latin *anonymus*, "what doesn't have a name or it is ignored, because it remains occult or unknown," according to the *Diccionario de Autoridades* of the Real Academia Española.

It designates not so much an absence as *the presence of an absence*, as Roland Barthes put it. The concept points out to the absence of a name *for the receiver of a message* (reader, viewer, critic, etc., reception instance which is constituent of "anonymity"), the absence of "signature," following Derrida. Anonymity is therefore closely linked to the forms of mediation, including writing. It implies the power to remain secret (without name) as author for a given audience. Seen from a discursive point of view, anonymity concerns associated with big data analysis are related to the generation of consistent narratives from massive and diverse amounts of data.

If we examine the concept from a textual perspective, we have to relate it to that of "author." When speaking of "anonymous author," we are already establishing a difference, taking up Foucault's terms, between the concepts of *proper name* (corresponding to the civilian, the physical, empirical individual; as Derrida pointed out: "the proper name belongs neither to language nor to the element of conceptual generality") and *name of the author* (situated in the plane of language, operating as a catalyst of textualities, as the lowest common denominator which agglutinates formal, thematic, or rhetoric specificities from different texts unified by a "signature"). If there is no author's name – "signature" – that is, if the text appears to be anonymous (from an "anonymous author"), this rubric loses the function of catalyst to become a generator of intransitive and individualized textualities, unable to be gathered into a unified *corpus*.

It is important to understand that the author we are talking about is not an empirical entity but a textual organizer. It does not necessarily match either the name of the empirical author, since a pseudonym could as well perform this function of textual organizer, because it keeps secret the proper name of the emitter. Foucault (1980: 114) clearly explained this point, in relation to the presence of names of authors in his work, and what meant for him, in theoretical terms, the *name of the author* (linked to what he calls "authorial function"), which some critics, dealing with his writings confused with the empirical subject (the "proper name"): "They ignored the task I had set myself: I had no intention of describing Buffon or Marx or of reproducing their statements or implicit meanings, but, simply stated, I wanted to locate the rules that formed a certain number of concepts and theoretical relationships in their works." Barthes (1977: 143) also alluded to the confusion between the proper name and the name of the author and its consequences: "Criticism still consist for the most part in saying that Baudelaire's work is the failure of Baudelaire the man." Jean-Paul Sartre (1947) was one of the most famous victims of that misunderstanding, reading Baudelaire's poems from a Freudian approach to the author's life and family traumas.

The words “Marx,” “Buffon,” or “Baudelaire” do not point to certain individuals with certain convictions, biographical circumstances, or specific styles, but to a set of textual regularities. In this sense, anonymity, in the context of the authorial function, points toward a *relational deficit*. To identify regularities, a minimum number of texts is required (a textual “family”) that permit to be gathered together through their “belonging” to the same “signature.” This socializing function of the nonanonymous author (becoming the centripetal force which allows that different texts live together) vanishes in the case of anonymous authors (or those which made use of different pseudonyms).

Let’s think, for example, of a classic novel, arrived to us under the rubric “anonymous,” whose author is, by chance, identified and given a name. From that moment on, the work will be “charged” with meanings from its incorporation to the collection of works signed by the author now identified. Similarly, the image we have today, for example, of a writer, politician, or philosopher, would be altered, i.e., reconstituted, if we found out that, avoiding his public authorial name, he had created texts whose ideological or aesthetic significance were inconsistent with his official production. Let us consider, for example, the eighteenth century fabulists (for instance, the French La Fontaine or the Spaniard Samaniego) whose official logic was one of a strict Christian morality, whereas some of their works, remained anonymous for a while and today attributed to them, could be placed within the realm of vulgar pornography.

In the textual Internet’s ecosystem, anonymity has become a hotspot for different reasons, and the issue is usually related to:

1. Power, referring to those who control the rules underlying Internet narratives (the programming that allows content display by users) and are able to take over the system (including hackers and similar characters; the *Anonymous* organization would be a good example, denomination included, and because of the
2. Extension of the above: anonymity as the ability to see without being seen. In this case, anonymity deepens the information gap (inequality of knowledge and asymmetric positions in the communication process). Those who are able to remain nameless are situated in a privileged position with respect to those who hold a name, because, among other things, they do not leave traces, they can not hardly be tracked, they have no “history,” therefore no past. It is no coincidence that when the “right” to anonymity is claimed by Internet users, it is formulated in terms of “right to digital oblivion.” Anonymous is the one that cannot be remembered. Anonymous is also the one who can see without being seen. In all cases, it implies an inequality of knowledge and manifests the oscillation between ignorance and knowledge.
3. Anonymity as a practice that permits some acts of speech go “without consequences” for the civilian person (for example, in the case of anonymous defamation or practiced under false names, or in the case of leaks), eluding potential sanctions (this brings us back to those authors forced to remain secret and hide their names in order to avoid being punished for their opinions, etc.). In this sense, anonymity may contribute to the advancement of knowledge by allowing the expression of certain individuals or groups whose opinions or actions won’t be accepted by the generality of the society (for example, the case of Mary Shelley, *Frankenstein’s* author, whose novel remained unsigned for a quarter of a century).
4. Anonymous is also the author who renounces the fame of the name, leaving the “auctoritas” to the text itself; the text itself would assume that role backing what is stated by the strength of its own argumentative power. In this sense,

paradox manifested on it of branded, i.e., publicized anonymity). Those who are able to determine the expressive and discursive modalities, subsequently fed by users’ activity, usually remain hidden or secret, i.e., anonymous.

anonymity is very well suited to permeate habits and customs. Anonymity also facilitates appropriation (for instance, in the case of plagiarism), reducing the risks of sanctions derived from the “private property of meaning” (which is what the signature incorporates to the text). As Georg Simmel wrote: “The more separate is a product from the subjective mental activity of its creator, the more it accommodates to an objective order, valid in itself, the more specific is its cultural significance, the more appropriate is to be included as a general means in the improvement and development of many individual souls (...) realisations that are objectified at great distance from the subject and to some extent lend ‘selflessly’ to be the seasons of mental development.”

5. Anonymity, in the unfolding discourse about mass media, has also been associated with the condition of massive and vicarious reception, made possible by the media, by the *anonymous masses*. In this sense, anonymity is associated with the indistinct, the lack of individuality, and the absence of the shaping and differentiating force of the name. As we see, extremes meet and connotations vary depending on the historical moment. Anonymity can both indicate a situation of powerlessness (referring, for example, to the masses) and a position of power (in the case, for example, of hackers or organizations or individuals who “watch” without being noticed the Internet traffic). Users “empowerment” through Internet and the stated passage from massive audiences to individualized users does not necessarily incorporate changes in authorial terms, because, as we have seen, we should not confuse the author’s name and the proper name. In the same way, authorial commitment and civilian commitment should be distinguished. In this sense, Walter Benjamin wrote in “The Author as Producer” (Benjamin 1998: 86): “For I hope to be able to show you that the concept of commitment, in the perfunctory form in which it generally occurs in the debate I have just mentioned, is a totally inadequate

instrument of political literary criticism. I should like to demonstrate to you that the tendency of a work of literature can be politically correct only if it is also correct in the literary sense. That means that the tendency which is politically correct includes a literary tendency. And let me add at once: this literary tendency, which is implicitly or explicitly included in every correct political tendency, this and nothing else makes up the quality of a work. It is because of this that the correct political tendency of a work extends also to its literary quality: because a political tendency which is correct comprises a literary tendency which is correct.” In this sense, all writing that makes a difference is anonymous from the point view of the “proper name.” This means that a consummated writing process inevitably leads to the loss of the proper name and designates the operation by which the individual who writes reaches anonymity and then *becomes an author* (anonymous or not).

6. Anonymity concerns related to big data should take into account the fact that those that “own” and sell data are not necessarily the same that generate those narratives, but in both cases the economic factor and the logic of profit optimization along with the implementation of control and surveillance programs are paramount. The “owners” are situated at the informational level, according to Shannon and Weaver’s notion of information. They established the paradigmatic context, the “menu,” within whose borders syntagmatic storytelling takes place through a process of data selection and processing. Users’ opinions and behaviors, tracked through different devices connected to the Internet, constitute the material of what we may call software driven storytelling. The fact that users’ information may be turned against them when used by the powers that be, which is considered one of the main privacy threats related to big data, reflects the fact that individuals, in the realm of mass media, have become “storytelling fodder,” which is probably the most extreme and oppressive

form of realism. Driven by institutionalized sources and power structures, the reading contract that lies beneath these narratives and the modes of existence of these discourses are structurally resilient to dissent.

In all these cases, the absence or the presence of the name of the author, and specifically anonymity, has to be considered as an institutionalized textual category consummated during the moment of reception/reading (emitting by itself does not produce anonymity, a process of reception is required; there is no anonymity without a “reading”), because it implies not so much a quality of the text as a “reading contract.” As Foucault (1980: 138) said, in a context of authorial anonymity, the questions to be asked will not be such as: “Who is the real author?,” “Have we proof of his authenticity and originality?,” “What has he revealed of his most profound self in his language?,” but questions of a very different kind: “What are the modes of existence of this discourse?,” “Where does it come from; how it is circulated; who controls it?,” “What placements are determined for possible subjects?,” “Who can fulfill these diverse functions of the subject?” It seems clear that the implications of considering one or another type of questions are not irrelevant, not only artistically or culturally, but also from a political perspective.

Further Reading

- Barthes, R. (1977). *The death of the author* (1967). In *Image, music, text*. London: Fontana Press.
- Benjamin, W. (1998). *The author as producer* (1934). In *Understanding Brecht*. London: Verso.
- Derrida, J. (1988). *Signature event context* (1971). In *Limited Inc*. Chicago: Northwestern University Press.
- Derrida, J. (2013). *Biodegradables* (1988). In *Signature Derrida*. Chicago: University of Chicago Press.
- Foucault, M. (1980). *What is an author* (1969). In *Language, counter-memory, practice: Selected essays and interviews*. New York: Cornell University Press.
- Sartre, J.-P. (1947). *Baudelaire*. Paris: Gallimard.
- Simmel, G. (1908). *Das Geheimnis und die geheime Gesellschaft*. In *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*. Leipzig: Duncker & Humblot.

Anonymization Techniques

Mick Smith¹ and Rajeev Agrawal²

¹North Carolina A&T State University,
Greensboro, NC, USA

²Information Technology Laboratory, US Army
Engineer Research and Development Center,
Vicksburg, MS, USA

Synonyms

[Anonymous data](#); [Data anonymization](#); [Data privacy](#); [De-Identification](#); [Personally identifiable information](#)

Introduction

Personal information is constantly being collected on individuals as they browse the internet or share data electronically. This collection of information has been further exacerbated with the emergence of the Internet of things and the connectivity of many electronic devices. As more data is disseminated into the world, interconnected patterns are created connecting one data record to the next. The massive data sets that are collected are of great value to businesses and data scientists alike. To properly protect the privacy of these individuals, it is necessary to de-identify or anonymize the data. In other words, personally identifiable information (PII) needs to be encrypted or altered so that a person’s sensitive data remains indiscernible to outside sources and readable to the pre-approved parties. Some popular anonymization techniques include noise addition, differential privacy, k-anonymity, l-diversity, and t-closeness.

The need for anonymizing data has come from the availability of data through big data. Cheaper storage, improved processing capabilities, and a greater diversity of analysis techniques have created an environment in which big data can thrive. This has allowed organizations to collect massive amounts of data on the customer/client base. This information in turn can then be subjected to a

variety of business intelligence applications so as to improve the efficiency of the collecting organization. For instance, a hospital can collect various patient health statistics over a series of visits. This information could include vital statistics measurements, family history, frequency of visits, test results, or any other health-related metric. All of this data could be analyzed to provide the patient with an improved plan of care and treatment, ultimately improving the patient’s overall health and the facilities ability to provide a diagnosis.

However, the benefits that can be realized from the analysis of massive amounts of data come with the responsibility of protecting the privacy of the entities whose data is collected. Before the data is released, or in some instances analyzed, the sensitive personal information needs to be altered. The challenge comes in deciding upon a method that can achieve anonymity and preserve the data integrity.

Noise Addition

The belief with noise addition is that by adding noise to data sets that the data becomes ambiguous and the individual subjects will not be identified. The noise refers to the skewing of an attribute so that it is displayed as a value within a range. For instance, instead of giving one static value for a person’s age, it could be adjusted ±2 years. If the subject’s age is displayed as 36, the observer would not know the exact value, only that the age may be between 34 and 38. The challenge with this technique comes in identifying the appropriate amount of noise. There needs to be enough to mask the true attribute value, while at the same time preserving the data mining relationships that exist within the dataset.

Differential Privacy

Differential privacy is similar to the noise addition technique in that the original data is altered slightly to prevent any de-identification. However, it is done in a manner that if a query is done on two databases that differ in only one row that the information contained in the missing

row is not discernable. Cynthia Dwork provides the following definition:

A randomized function K gives ε-differential privacy if for all data sets D₁ and D₂ differing on at most one element, and all S ⊆ Range(K),

$$Pr[K(D_1) \in S] \leq exp(\epsilon) \times Pr[K(D_2) \in S]$$

As an example think of a database containing the incomes of 75 people in a neighborhood and the average income is \$75,000. If one person were to leave the neighborhood and the average income dropped to \$74,000, it would be easy to identify the income of the departing individual. To overcome this, it would be necessary to apply minimum noise so that the average income before and after would not be representative of the change. At the same time, the computational integrity of the data is maintained. The amount of noise and whether an exponential or Laplacian mechanism is used is still subject to ongoing research/discussion.

K-Anonymity

In the k-anonymity algorithm, two common methods for anonymizing data are suppression and generalization. By using suppression, the values of categorical variable, such as name, are removed entirely from the data set. With generalization quantitative variables, such as age or height, are replaced with a range. This in turn makes each record in a data set indistinguishable from at least k–1 other records. One of the major drawbacks to k-anonymity is that it may be possible to infer identity if certain characteristics are already known. As a simple example consider a data set that contains credit decisions from a bank (Table 1). The names have

Anonymization Techniques, Table 1 K-anonymity credit example

Age	Gender	Zip	Credit decision
18–25	M	149**	Yes
18–25	M	148**	No
32–39	F	149**	Yes
40–47	M	149**	Yes
25–32	F	148**	No
32–39	M	149**	Yes

been omitted, the age categorized, and the last two digits of the zip code have been removed.

This obvious example is for the purposes of demonstrating the weakness of a potential homogeneity attack in k -anonymity. In this case, if it was known that a 23-year-old man living in 14,999 was in this data set, the credit decision information for that particular individual could be inferred.

L-Diversity

L-diversity can be viewed as an extension to k -anonymity in which the goal is to anonymize specific sensitive values of a data record. For instance, in the previous example, the sensitive information would be the credit decision. As with k -anonymity generalization and suppression techniques are used to mask the true values of the target variable. The authors of the l-diversity principle, Ashwin Machanavajhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkatasubramniam, define it as follows:

A q^ -block is l -diverse if it contains at least l well-represented values for the sensitive attribute S .
A table is l -diverse if every q^* -block is l -diverse.*

The concept of well-represented has been defined in three possible methods: distinct l -diversity, entropy l -diversity, and recursive (c , l)-diversity. A criticism of the l -diversity model is that it does not hold up well when the sensitive value has a minimal number of states. As an example, consider the credit decision table from above. If that table were extended to include 1000 records and 999 of them had a decision of “yes,” then l -diversity would not be able to provide sufficient equivalence classes.

T-Closeness

Continuing with the refinement of de-identification techniques, t -closeness is an extension of l -diversity. The goal of t -closeness is to create equivalence classes that approximate the original

distribution of the attributes in the initial database. Privacy can be considered a measure of information gain. T -Closeness takes this characteristic into consideration by assessing an observer’s prior and posterior belief about the content of a data set as well as the influence of the sensitivity attribute. As with l -diversity, this approach hides the sensitive values within a data set while maintaining association through “closeness.” The algorithm uses a distance metric known as the Earth Mover Distance to measure the level of closeness. This takes into consideration the semantic interrelatedness of the attribute values. However, it should be noted that the distance metric may differ depending on the data types. This includes the following distance measures: numerical, equal, and hierarchical.

Conclusion

To be effective each anonymization technique should prevent against the following risks: singling out, linkability, and inference. Singling out is the process of isolating data that could identify an individual. Linkability occurs when two or more records in a data set can be linked to either an individual or grouping of individuals. Finally inference is the ability to determine the value of the anonymized data through the values of other elements within the set. An anonymization approach that can mitigate these risks should be considered robust and will reduce the possibility of re-identification. Each of the techniques presented address each of these risks differently. The following table outlines their respective performance (Table 2):

For instance, unlike k -anonymity, l -diversity, and t -closeness are not subject to inference attacks that utilize the homogeneity or background knowledge of the data set. Similarly, the three generalization techniques (k -anonymity, l -diversity, and t -closeness), all present differing levels of association that can be made due to the clustering nature of each approach.

As with any aspect of data collection, sharing, publishing, and marketing, there is the potential

Anonymization Techniques, Table 2 Anonymization algorithm comparison

Technique	Singling out	Linkability	Inference
Noise addition	At risk	Possibly	Possibly
K-anonymity	Not at risk	At risk	At risk
L-diversity	Not at risk	At risk	Possibly
T-closeness	Not at risk	At risk	Possibly
Differential privacy	Possibly	Possibly	Possibly

for malicious activity. However, the benefits that can be achieved from the potential analysis of such data cannot be overlooked. Therefore, it is extremely important to mitigate such risks through the use of effective de-identification techniques so as to protect sensitive personal information. As the amount of data becomes more abundant and accessible, there is an increased importance to continuously modify and refine existing anonymization techniques.

Further Reading

- Dwork, C. (2006). Differential privacy. In *Automata, languages and programming*. Berlin: Springer.
- Li, Ninghui, et al. (2007). *t-Closeness: Privacy beyond k-anonymity and l-diversity*. IEEE 23rd International Conference on Data Engineering, 7.
- Machanavajjhala, A., et al. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), Article 3, 1–12.
- Sweeney, L. (2002). k-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5).
- The European Parliament and of the Council Working Party. (2014). Opinion 05/2014 on anonymisation techniques. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf. Retrieved on 29 Dec 2014.

Anonymous Data

► Anonymization Techniques

Anthropology

Marcienne Martin

Laboratoire ORACLE [Observatoire Réunionnais des Arts, des Civilisations et des Littératures dans leur Environnement] Université de la Réunion Saint-Denis France, Montpellier, France

Irrespective of the medium applied, information compiles data relating to a given study object. This is the case with anthropology. Indeed, diversity translated through language, culture, or social structure is an important source of information. Concerning the study of human beings, answers have differed according to the different epochs and cultures. Anthropology as a field of scientific research began in the nineteenth century. It derived from anthropometry, a science dedicated to the dimensional particularities of human being. Buffon with the study: *Traité des variations de l'espèce humaine* (Study on the Variation of the Human Species) (1749) and Pierre-Paul Broca, the founder of the Society of Anthropology of Paris (1859), are considered rance's forerunners of this science. In the era of the Internet, data has become accessible to nearly anyone wishing to consult them. In the free encyclopedia Wikipedia, there are over 1,864,000 articles in the French language. As for the eleven thematic portals: art, geography, history, leisure, medicine, politics, religion, science, society, sport, technology, they subsume 1,636 portals, always in the French language. Anthropology is one of the entries of this encyclopedia. Anthropology refers to the science dedicated to the study of the human as a whole, either at the physical level, as it belongs to the animal world, or in the context of both its environment and history when analyzed from the perspective of different human groups who have been observed. From an etymological point of view, the term “anthropology” stems from the Greek “Anthropos,” which contrasts the human to the gods; moreover, the Greek word “logos” refers to “science, speech.” The anthropologist, a specialist in this field of research, is written in the Greek

language as follows: ἀνθρώπος ὁ λόγιος. Other related sciences, such as anthropology, sociology, etc., also study the *Homo sapiens*, but in a particular context (ethnicity, sociocultural substrate ...).

In general, a human being needs references. *Homo sapiens* always responds to questioning as it strange, with no less singular explanation, but sometimes validated by the phenomenon of beliefs or hypothesis made by them, perhaps depending on the evolution of technology through verifiable hypothesis (dark matter, dark energy ...). These interrogations are at the origin of the creation of mythologies, of religions, and of philosophies. So to answer the question of the origin and the meaning of natural phenomena, such as thunderstorms, volcanoes, and storms, diverse beliefs have attributed these creations to the deities, sometimes in response to human behaviors considered as negative; hence, these deities have developed these natural disasters. When these phenomena were understood scientifically, the responses related to these beliefs disappeared. Why we live in order to finally die is a question that has not been answered satisfactorily yet, except through many religions where beliefs and their responses are posed as postulates.

Nourished by the richness of imagination in humans, philosophy is a mode of thinking which tries to provide responses to various questions concerning the meaning of life, to the various human behaviors and to their regulations (moral principles), to the death, to the existence or to the inexistence of an “architect” at the origin of the world of matter. Concerning the language as a method of transmission, regardless of the tool used, the thought is based on complex phenomena. The understanding of the objects in the world induces different forms of reasoning, such as logical reasoning and analogical reasoning. Some types of reasoning include a discourse *de facto*, regardless of the type of reasoning (concessive reasoning, dialectic reasoning, by *reductio ad absurdum*) and whatever its form (oral, written). In contrast, both the inductive and deductive reasoning type, even if they are integrated in the processes of discursive types, are correlated to the description of the objects of the world and to their place in the human paradigm. As for logical

reasoning, the observed object in its contextual relationships is taken into consideration and the concluding speech serves as the culmination of the progress of thought. There is also another type of reasoning; it is analogical reasoning. In this cognitive process, an unknown object is put into relation, or where some of its given parameters are incomprehensible, but have some resemblance with something known, at least according to the observer's perception.

Between differentiation and analogy, Human has built different paradigms one of which was incorporated divine entities. In contrast, some are elaborated from elements of objects in the world, e.g., the Centaur, half horse, half human. Others use the imaginary as a whole, which corresponds to the rewriting the objects of the world in a recomposition *ad infinitum*. For the Greeks, the principle of anthropomorphization of phenomena, as still unknown or objects whose origin was still unexplainable (Gaia the Earth), has been extended to some objects such as the Night, the Darkness, the Death, the Sleep, the Love, and the Desire (Gomperz 1908).

The major revolution that has given a new orientation to the study of human beings, as a species which belongs to the world of the living, was the hypothesis made by the English naturalist Darwin in 1859 about the origin of species, their variability, their adaptation to their environment, and their evolution. If the laws of heredity were not yet known at this period, it is the Czech-German monk and botanist Gregor Johann Mendel, who developed the three laws of heredity, known as Mendel's laws, in 1866 after 10 years of study on hybridization in plants. In the journal *Nature*, in a paper published in 1953, the researchers James Watson and Francis Crick demonstrated the existence of the double helical structure of DNA. According to Crick, the analysis of the human genome is an extraordinary approach which is applied to the human being with respect to both its development and physiology with the benefits that it can give such in the medical field.

In addition, other researchers have worked on the phenomenon of entropy and negative entropy at the origin of the transformation of units composing the living world. Schrödinger (1993) has

put in relation entropy, energy exchanges, probability, and the order of life. Moreover, Monod (1960) evokes the emergence of life and, implicitly, that of the human as pure chance. Prigogine (1967) reassesses the question of asking about the nature of the living world and, therefore, the human, based on the main principles of physics as well as those of thermodynamics; the first principle affirms the conservation of energy by all systems and the second principle (principle of order Boltzmann) holds that an isolated system evolves spontaneously toward a state of balance which corresponds to maximum entropy.

Bio sociology is a particular approach applied to the world of the living. So research of the ethologist Jaisson (1993) addresses the social insects, including ants. This author shows that there is a kind of programming that is the cause of the behavior of the species *Formicus* (ant) belonging to the order of *Hymenoptera*. This study is similar to that done by Dawkins (2003), an ethologist who supported the evolutionary theory of Darwin but posits that natural selection would be initiated by the gene through an existing program, not by the species. This observation puts the innate and the acquired as parameters of human culture into question.

These studies have opened an exploratory field to the evolutionary biologists, such as Diamond (1992) who showed the phylogenetic similarity between the pygmy chimpanzee of Zaire and the common chimpanzee from Africa and *Homo sapiens*. These results are based on the molecular genetic studies which have shown that we share over 98% of our genetic program with these primates. The 2% which make the difference are somehow the magical openings which allow us, in our role as human beings, to access to the understanding of the universe in which we live. This understanding is correlated to the awareness of existence and its manifestation through discourse and language. Leroi-Gourhan (1964) has stipulated that two technical centers in many vertebrates result in anthropoids for the formation of two functional pairs (hand tool and face-language). The emergence of the graphic symbol at the end of the reign of the Paleanthropien entails forging new relationships between two operative

poles. In this new relationship, the vision holds the greatest place in the pairs: face-reading and hand-graphy.

If we continue the analysis between hominids and other members of the living world, we find that the observation of the environment made by the whole of the living world is intended to protect the species and the diffusion relative to their survival. These operating modes involve the instinct, either a biological determinism which, in a particular situation, responds to a special behavior and refers to a basic programming more or less adaptable depending on the species. Moreover, whether breeding rituals, love rituals, or the answers to a situation of aggression, behaviors will be similar to a member of the species to another; indeed, the survival of the species takes priority over that of the member of the species. Among the large primate, these answers become more appropriate and they open the field of a form of creativity. In an article on chimpanzees, Servais (1993) states that these primates do not have any language; they communicate only by manipulating their behavior. They are able, for the most intelligent of them, to associate, to form coalitions, to conclude pacts, or to have access to a form of concept thought. They have forms of "protoculture"; the most famous is without doubt the habit of washing one's potatoes in some groups of Japanese macaques, but they have no cultural production; they have a typical social organization in relation with their species, but they have no written or oral laws. This punctual creativity among the great simians has grown exponentially in humans; it is that form of creativity which opened the field of imaginary.

If the questioning concerning the innate and the acquired has been the subject of various experiences, the study of diverse ethnic groups demonstrates that through culture the adaptation of the human is highly diversified. This is due to the genealogical chain which shows the phenomenon of nomination, which, in turn, is in resonance with the construction of individual identity. The anthropologist and ethnologist Levi-Strauss (1962) evokes different modes of naming in use in ethnic groups like the Penan of Borneo as, e.g., the tecknonym meaning "father of such a" or

“mother of such a” or the necronym which expresses the family relationship existing with a deceased relative and the individually named. Emperaire (1955), a researcher at the Musée de L’homme in Paris, gives the example of the Alakalufs, an ethnic group living in Tierra del Fuego, which does not name the newborn at birth; the children do not receive a name; it is only when they begin to talk and walk that the father chooses one. Other systems of genealogical chain, such as those designated as “rope” and which correspond to a link which groups a man, his daughter and the son of his daughter or a wife, son and daughters of his son (Mead 1963).

Cultural identity is articulated around specific values belonging to a particular society and defining it; they have more or less strong emotional connotations; thus a taboo object should be lived out as a territory to not transgress, because the threat of various sanctions exist including the death of the transgressor. The anthropologist Mead (1963) has exemplified this phenomenon by studying the ethnic group of the Arapesh, who were living in the Torricelli Mountains in New Guinea at the time when the author was studying their way of life (1948). The territories of the male group and the territories of the female group were separated by territories marked as taboos; concerning the flute, an object belonging to the male group, for the female group, this object was prohibited. The semantic content of certain lexical items may differ from one ethnic group to another, and even sometimes may become an antinomy. Mead cites the Arapesh and the Mundugumor, two ethnic groups that have developed their identity through entirely different moral values and behaviors. Thus, Arapesh society considers each member as sweet and helpful and wants to avoid violence. In contrast, in the ethnic group of the Mundugumor, their values are the antonyms of the ethnic group of the Arapesh.

As for big data, the implications can vary from one culture to another: highlighting history, traditions, social structure, the official language, etc. The addition of data by Internet users, according to their desires and their competencies, contribute to the development of the free encyclopedia Wikipedia. Each user can contribute by

adding an article or correcting it. The user of the Internet then plays the role of a contributor, i.e., writer and corrector; he or she can also report false information appearing in the context of articles written by other Internet users. This multiple role is equivalent to what the philosopher and sociologist Pierre Levy calls “collective intelligence,” namely, the interactions of the cognitive abilities of the members of a given group enabling them to participate in a common project.

Within the framework of more specialized research domains, many university websites on the Internet offer books and magazines, which cannot always be consulted free of charge or without prior registration. Today’s access to big data differs from the period before the arrival of digital technologies, where only the university libraries were able to meet the demands of students and researchers, both in terms of the availability of works and of their varieties.

Access to online knowledge has exponentially multiplied the opportunity for anyone to improve their knowledge within a given field of study.

Further Reading

- Dawkins, R. (2003). *Le gène égoïste*. Paris: Éditions Odile Jacob.
- De Buffon, G.-L. (1749–1789). *Histoire Naturelle générale et particulière: avec la description du Cabinet du Roy*, par Buffon et Daubenton. Version en ligne au format texte. <http://www.buffon.cnrs.fr/index.php?lang=fr#hn>.
- Diamond, J. (1992). *Le troisième singe – Essai sur l’évolution et l’avenir de l’animal humain*. Paris: Gallimard.
- Emperaire, J. (1955). *Les nomades de la mer*. Paris: Gallimard.
- Gomperz, T. (1908). *Griechische Denker: eine Geschichte der antiken Philosophie. Les penseurs de la Grèce: histoire de la philosophie antique* (Vol. 1). Lausanne: Payot. <http://catalogue.bnf.fr/ark:/12148/cb30521143f>.
- Jaisson, P. (1993). *La fourmi et le sociobiologiste*. Paris: Éditions Odile Jacob.
- Leroi-Gourhan, A. (1964). *Le Geste et la Parole, première partie: Technique et langage*. Paris: Albin Michel.
- Lévi-Strauss, C. (1962). *La pensée sauvage*. Paris: Plon.
- Lévy, P. (1997). *L’intelligence collective – Pour une anthropologie du cyberspace*. Paris: Éditions La Découverte Poche.
- Mead, M. (1963). *Mœurs et sexualité en Océanie – Sex and temperament in three primitive societies*. Paris: Plon.
- Monod, J. (1960). *Le hasard et la nécessité*. Paris: Seuil.

- Prigogine, I. (1967). *Introduction to thermodynamics of irreversible processes*. New York: John Wiley Interscience.
- Roger J. (2006). *Buffon*. Paris: Fayard.
- Schrödinger, E. (1993). *Qu'est-ce que la vie ? : De la physique à la biologie*. Paris: Seuil.
- Servais, V. (1993). Les chimpanzés: un modèle animal de la relation clientélaire. *Terrain*, 21. <https://doi.org/10.4000/terrain.3073>. <http://terrain.revues.org/3073>.
- Wiener, N. (1948). *Cybernetics or control and communication in the animal and the machine*. Cambridge, MA: MIT Press.

Antiquities Trade, Illicit

Layla Hashemi and Louise Shelley
Terrorism, Transnational Crime, and Corruption
Center, George Mason University, Fairfax, VA,
USA

The cyber environment and electronic commerce have democratized the antiquities trade. Previously, the antiquities trade consisted of niche networks of those associated with galleries and auction houses. While the trade was transnational, there were high barriers to entry preventing the average person from becoming involved. Today's antiquities market has been democratized by the internet and online platforms that allow for the free and often open sale of cultural property and the availability of ancient goods, particularly coins, often at affordable prices. Therefore, what was once a high-end and exclusive commodity is now available to a much broader and less sophisticated customer base. Identifying the actors behind this trade and understanding the extent and profits of this trade requires large-scale data analytics.

Many vendors use open web platforms such as VCoins, Etsy, eBay, and other marketplaces to advertise and sell their products. These sales are possible because Section 230 of the US Communications Decency Act releases websites from responsibility for the content posted on their platforms, allowing criminals to conduct their business online with near total impunity. Recent analysis in 2020 revealed that around 2.5 million

ancient coins are being offered on eBay annually with actual sales estimated at \$26–59 million (Wartenberg and Brederova 2021). This massive supply of coins readily available to customers could only be achieved through the extensive looting of archaeological sites in the Middle East.

With a low barrier to entry, online marketplaces allow any individual interested in cultural heritage to collect information and even purchase the items at affordable costs. Because the antiquities trade is a gray trade where there is a mixing of licit and illicit goods, these transactions are often completed with impunity. Therefore, sellers currently do not need to trade on the dark web as they seek to reach the largest number of customers and are not deterred by the actions of law enforcement who rarely act against online sellers of antiquities (Brodie 2017).

Social media platforms are also often used in the antiquities trade. Platforms such as Facebook allow traffickers to reach a large audience with a casual interest in antiquities, thus normalizing the idea of looting for profit (Sargent et al. 2020). These online venues range from private Facebook groups to fora used to discuss the authenticity and value of specific items (Al-Azm and Paul 2018, 2019). Moreover, once the initial contact between the seller and the buyer is made, the trade often moves to encrypted channels, protecting both the seller and the purchaser from detection. While Facebook recently announced a ban on the sale of historical artifacts, there are still strong indications that sales have not ceased on the platform (Al-Azm and Paul 2019).

Detection of participants in the trade is difficult because of the large volume of relatively small transactions. One solution is the blending of manual and automated computational methods, such as machine learning and social network analysis, to efficiently process data and identify leads and gather investigative evidence. Using large sets of interoperable data, investigative leads can be supplemented with financial, transport, and other data to examine entire supply chains from source through transit to destination countries. Financial investigations of the transfer of digital assets allow for the mapping of

transnational transactions through the analysis of payment processing. Only with techniques using sophisticated data analytics will it be possible for investigators to address these crimes. At present, creative and innovative criminal actors frustrate the ability of governments to disrupt this pervasive online criminal activity causing irreparable damage to the international community's cultural heritage.

Cross-References

- [Persistent Identifiers \(PIDs\) for Cultural Heritage](#)
- [Transnational Crime](#)

Further Reading

- Al-Azm, A., & Paul, K. A. (2018). How Facebook made it easier than ever to traffic middle eastern antiquities. <https://www.worldpoliticsreview.com/insights/25532/how-facebook-made-it-easier-than-ever-to-traffic-middle-eastern-antiquities>. Accessed 22 Dec 2020.
- Al-Azm, A., & Paul, K. A. (2019). Facebook's black market in antiquities: Trafficking, terrorism, and war crimes. <http://atharproject.org/wp-content/uploads/2019/06/ATHAR-FB-Report-June-2019-final.pdf>. Accessed 22 Dec 2020.
- Brodie, N. (2017). How to control the Internet market in antiquities? The need for regulation and monitoring. Antiquities coalition, policy brief no. 3.
- Brodie, N. (2019). Final report: Countering looting of antiquities in Syria and Iraq. <https://traccc.gmu.edu/sites/default/files/Final-TraCCC-CTP-CTAQM-17-006-Report-Jan-7-2019.pdf>.
- Brodie, N., & Sabrine, I. (2018). The illegal excavation and trade of Syrian cultural objects: A view from the ground. *Journal of Field Archaeology*, 43(1), 74–84. <https://doi.org/10.1080/00934690.2017.1410919>.
- Sargent, M., et al. (2020). Tracking and disrupting the illicit antiquities trade with open source data. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2706.html. Accessed 22 Dec 2020.
- Wartenberg, U., & Brederova, B. (2021). Plenitudinous: An analysis of ancient coin sales on eBay in. In L. Hashemi & L. Shelley (Eds.), *Antiquities smuggling: In the real and the virtual world*. Abingdon/New York: Routledge.
- Westcott, T. (2020). Destruction or theft: Islamic state, Iraqi antiquities and organized crime. <https://globalinitiative.net/wp-content/uploads/2020/03/Destruction-or-theft-Islamic-State-Iraqi-antiquities-and-organized-crime.pdf>. Accessed 22 Dec 2020.

Apple

R. Bruce Anderson^{1,2} and Kassandra Galvez²

¹Earth & Environment, Boston University,
Boston, MA, USA

²Florida Southern College, Lakeland, FL, USA

In 1984, Steve Jobs and Steve Wozniak started a business in a garage. This business went on to change the global dynamic of computers as it is known today. Apple Inc. all started because Jobs and Wozniak wanted the machines to be smaller, cheaper, intuitive, and accessible to everyday consumers, but more important, user-friendly. Over the past 30 years, Apple Inc. has transformed this simple idea into a multi-billion dollar industry that includes: laptops, desktops, tablets, music players, and so much more. The innovative style and hard-wired simplicity of Apple's approach has proven to be a sustained leader for computer design.

After some successes and failures, Apple Inc. created one of the most revolutionary programs to date: iTunes. In 2001, Apple released the iPod – a portal music player. The iPod allowed consumers to place music files into the iPod for music “on the go”; however, instead of obtaining the music files from CD's, you would obtain them online, via a proprietary website “iTunes”. iTunes media players systematically changed the way music is played and purchased. Consumers could now purchase digital copies of albums instead of “hardcopy” discs. This affordable way of purchasing music impacted the music industry in an enormous way. Its impact is unparalleled in terms of how the music industry profits off music sales and how new artists have been able to break through. The iTunes impact, however, reaches far beyond the music industry. Podcasts have impacted the way we can access educational material, for example. Music and media is easily accessible from anywhere in the world via iTunes.

While Macs and MacBooks are one of their most profitable items, the iPhone, created in 2007, really changed the industry. When Steve Jobs first introduced the iPhone, it was so different from

other devices because it was a music player, a phone, and an internet device all in one. With the iPhone's touchscreen and other unique features, companies like Nokia and Blackberry were left in the dust which resulted in many companies changing their phone devices structural model to have similar features to the iPhone.

In 2010, Apple Inc. released a tablet known as the iPad. This multi-touch tablet features a camera, a music player, internet access, and applications. Additionally, the iPad has GPS functions, email, and video recording software. The iPad transformed the consumer image of having a laptop. Instead of carrying around a heavy laptop, consumers have the option of purchasing a lightweight tablet that has the same features as a laptop.

On top of these unique products, Apple Inc. utilizes the iOS operating system. iOS uses touch-based instructions such as: swipes, taps, and pinches. These various instructions provide specific definitions within the iOS operating system. Since its debut in 2007, iOS software has transformed the nature of phone technology. With its yearly updates, the iOS software has added more distinctive features from Siri to the Game Center.

Siri is a personal assistant and knowledge navigator that is integrated into the iOS software. Siri responds to spoken commands and allows the user to have constant hands-free phone access. With the sound of your voice and a touch of a button, Siri has full access to the user's phone. Siri can perform tasks such as calling, texting, searching the web, finding directions, and answering general questions. With the latest iOS update in the Fall of 2013, Siri expanded its knowledge and is now able to support websites such as: Bing, Twitter, and Wikipedia. Additionally, Siri's voice was upgraded to sound more man than machine.

Apple's entry into the world of Big Data was late – but a link-up with IBM has helped a great deal in examining how users actually use their products. The Apple Watch, which made its debut in 2015, is able to gather data of a personal nature that lifts usage data to a new level. The success of the applications associated with the Apple Watch and the continuing development of data-gathering apps for the iPad and other Apple

products has brought Apple at least in line with other corporations using Big Data as a source for consumer reflection information.

Apple Inc.'s advanced technology has transformed the computer industry and has made it one of the most coveted industries. In 2014, Apple Inc. is the world's second-largest information technology company by revenue after Samsung and the world's third-largest mobile phone maker after Samsung and Nokia. One of the ways Apple Inc. has become such an empire is with its retail stores. As of August 2014, Apple has 434 retail stores in 16 countries and an online store available in 43 countries. The Apple Store is a chain of retail stores owned and operated by Apple Inc., which deals with computers and other various consumer electronics. These Apple Stores sell items from iPhone, iPads, MacBooks, iPods, to third party accessories.

The access of third party applications to Apple hardware – though still relatively closely controlled – make it possible for Apple to utilize Big Data in ways not anticipated in early development of what was essentially a sealed system.

Since its origin, one of Apple Inc.'s goals was to make computer accessible to everyday people. Apple Inc. accomplished this goal by partnering with President Barack Obama's ConnectED initiative. In June 2013, President Obama announced the ConnectED initiative, designed to enrich K-12 education for every student in America. ConnectED empowers teachers with the best technology and the training to make the most of it and empowers students through individualized learning and rich, digital content. President Obama's mission is to prepare America's students with the skills they need to get good jobs and compete with other countries which rely increasingly on interactive, personalized learning experiences driven by new technology. President Obama states that fewer than 30% of America's schools have the broadband they need to teach using today's technology. However, under ConnectED, 99% of American students will have access to next-generation broadband by 2017. That connectivity will help transform the classroom experience for all students, regardless of income. President Obama has also directed the

federal government to make better use of existing funds to get Internet connectivity and educational technology into classrooms, and into the hands of teachers trained on its advantages, and he called on businesses, states, districts, schools, and communities to support this vision, which requires no congressional action.

Furthermore, in 2011, Apple Inc. partnered with “Teach for American,” a program that trains recent graduates from some of America’s most prestigious universities to teach in the meanest and most dangerous schools throughout the nation and donated over 9000 first generation iPads to teachers that work in impoverished and dangerous schools. These donated iPads came from customers who donated to Apple’s public service program during the iPad 2 launch. These 9000 first generation iPads were distributed to teachers in 38 states.

In addition to President Obama’s ConnectED initiative, Apple Inc. has also provided students and educators with special discounts which enable these devices to be much more accessible and affordable. During the months of June to August, Apple Inc. bestows up to \$200.00 in savings on specific Apple products such as MacBooks and iPads to students and educators. The only requirement to receive this discount is student identification or an educator’s identification. Once the proper verification is shown, students and educators receive the discount in addition to up to \$100.00 in iTunes and/or App Store Credit.

Apple Inc. caters to students not only with these discounts but also with various other education resources. Some of these education resources include: iBooks and iTunes U. iBook is an online store through Apple Inc. that allows the user to purchase electronic books. These electronic books are linked to your account and allow you access wherever you may be with the device. iBooks include materials from novels, travel guides, and textbooks. iTunes U is program for professors that enables students to access course materials, track assignments, and organize notes. Additionally, students can create discussions posts for that specific class, add material from outside sources, and generate a more specialized course. iTunes U not only offers elements for courses but it also

provides other education facets such as: interviews, journals, self-taught books, and more.

With the variety of products that consumers can buy through Apple, iCloud has proven to be a distinct source to store all users’ information. iCloud is cloud storage and computing service that was launched in 2011. iCloud allows users to store data such as music and other iOS applications on computer servers for download to multiple devices. Additionally, iCloud is a data syncing center for email, contacts, calendars, bookmarks, notes, reminders, documents, photos and other data. As such, much of the data in a “macro” setting is available to developers as aggregated material. However, in order to use these functions, users must create an Apple ID. An Apple ID is the email you use as a login for every Apple function such as buying songs on iTunes and purchasing apps from the App Store. By choosing to use the same Apple ID, costumers have the ability to keep all their data in one location. When costumer set up their iPhone, iPad, or iPod touch, they can use the same Apple ID for iCloud services and purchases on the iTunes Store, App Store, and iBooks Store. On top of that, users can set up their credit card and billing information through the Apple ID. This Apple ID allows users to have full access to any purchases on the go through iCloud. How much data is shared by Apple through the cloud remains something of a mystery. While individual user information is unlikely to be available, Big Data – data at the aggregate level measuring consumer usage – almost certainly is.

iCloud allows users to back up the setting and date on any iOS devices. This date includes photos and videos, device settings, app data, messages, ringtones, and visual voicemails. These iCloud backups occur daily the minute one of the consumers’ iOS devices is connected to Wi-Fi and a power source. Additionally, iCloud backs up contact information, email accounts, and calendars. Once the data is backed up, customers will have all the same information on every single iOS device. For example, if a user has an iPad, an iPhone, and a MacBook and starts adding schedules to their iPhone calendar, the minute the backup begins, he or she will be able to access that same calendar with the new schedule on his or

her iPad and MacBook. Again, Apple provides users with this unique connectivity by the use of an Apple ID and iCloud. When signing up for iCloud, users automatically get 5 gigabytes of free storage. In order to access more gigabytes, users can either go to their iCloud account to delete data or users can upgrade. When upgrading, users have three choices: a 10 gigabyte upgrade, a 20 gigabyte upgrade, and a 50 gigabyte upgrade. These three choices are priced at \$20.00, \$40.00, and \$100.00, respectively. These storage upgrades are billed annually.

The last two unique features that iCloud provides is Find my iPhone and iCloud Keychain. Find my iPhone allows users to track the location of their iOS device or Mac. By accessing this feature, users can see the device's approximate location on a map, display a message and/or play a sound on the device, change the password on the device, and remotely erase its contents. In recent upgrades, iOS 6 introduced Lost Mode, which is a new feature that allows users to mark a device as "lost," making it easier to protect and find. The feature also allows someone that finds the user's lost iPhone to call the user directly without unlocking it. This feature has proved to be useful in situations where devices are stolen. Since the release of this application in 2010, similar phone finders have become available for other "smart" phones.

The iCloud Keychain functions as a secure database that allows information including a user's website login passwords, Wi-Fi network passwords, credit/debit card management, and other account data, to be securely stored for quick access and auto-fill on webpages and elsewhere when the user needs instant access to them. Once passwords are in the iCloud Keychain, they can be accessed on all devices connected to the Apple ID. Additionally, to view the running list of passwords, credit and debit card information, and other account data, the user must put in a separate password in order to see the list of secure data. iCloud also has a security function. If users enter an incorrect iCloud Security Code too many times when using iCloud Keychain, the users iCloud Keychain is disabled on that device, the keychain in the cloud is deleted, and the user will receive

one of these alerts: "Security Code Incorrectly Entered Too Many Times. Approve this iPhone from one of your other devices using iCloud Keychain. If no devices are available, reset iCloud Keychain." or "Your iCloud Security Code has been entered too many times. Approve this Mac from one of your other devices using iCloud Keychain. If no devices are available, reset iCloud Keychain."

In 2013, Apple Inc. released its most innovative feature yet: Touch ID. Touch ID is a finger print scanner which doubles as a password protection on the iPhone 5 s, which is the latest version of the iPhone. The reason for making Touch ID is because more than 50% of users do not use a passcode: with Touch ID, creating and using a passcode is seamless. When accessing the iPhone 5S, users register every single finger into the system. By allowing this registration, users are able to unlock their iPhones with any finger. To unlock the iPhone, users simply place their finger on the home button; the Touch ID sensor reads the finger print and immediately unlocks the iPhone. Touch ID is not only for passwords but it also authorizes purchases onto your Apple ID such as: iTunes, iBooks, and the App store. On announcing this feature, Apple stated that Touch ID doesn't store any images of your fingerprint. It stores only a mathematical representation of your fingerprint. The iPhone 5S also includes a new advanced security architecture called the Secure Enclave within the A7 chip, which was developed to protect passcode and fingerprint data, which means that the Fingerprint data is encrypted and protected with a key available only to the Secure Enclave. Therefore, your fingerprint is never accessed by iOS or other apps, never stored on Apple servers, and never backed up to iCloud or anywhere else.

As secure as Apple's data likely is to outside investigators or hackers, it seems very likely that sampling at the Big Data level is constant.

Cross-References

- [Business Intelligence](#)
- [Cell Phone Data](#)

- [Cloud Computing](#)
- [Cloud Services](#)
- [Data Storage](#)
- [Education and Training](#)
- [Voice Data](#)

Further Reading

- Atkins, R. Top stock in news: Apple (NASDAQ AAPL) plans on building the world's biggest retail store. *Tech News Analysis*. 21 Aug. 2014. Web. 26 Aug. 2014.
- ConnectED Initiative. *The White House*. The White House, 1 Jan. 2014. Web. 26 Aug. 2014.
- Happy Birthday, Mac. *Apple*. Apple, 1 Jan. 2014. Web. 26 Aug. 2014.
- Heath, A. Apple donates 9,000 iPads to teachers working in impoverished schools. *Cult of Mac*. 20 Sept. 2011. Web. 26 Aug. 2014.
- iCloud: About iCloud security code alert messages. *Apple Support*. Apple, 20 Oct. 2013. Web. 26 Aug. 2014.
- iPhone 5s: About touch ID security. *Apple Support*. Apple, 28 Mar. 2014. Web. 26 Aug. 2014.
- Marshal, G. 10 Ways Apple Changed the World. *TechRadar*. 10 Mar. 2013. Web. 26 Aug. 2014.

Archaeology

Stuart Dunn
 Department of Digital Humanities, King's
 College London, London, UK

Introduction

In one sense, archaeology deals with the biggest dataset of all: the entire material record of human history, from the earliest human origins c. 2.2 million years Before Present (BP) to the present day. However this dataset is, by its nature, incomplete, fragmentary, and dispersed. Archaeology therefore brings a very particular kind of challenge to the concept of big data. Rather than real-time analyses of the shifting digital landscape of data produced by the day to day transactions of millions of people and billions of devices, approaches to big data in archaeology refer to the sifting and reverse-engineering of masses of data derived from both primary and

secondary investigation into the history of material culture.

Big Data and the Archaeological Research Cycle

Whether derived from excavation, post-excavation analysis, experimentation, or simulation, archaeologists have only tiny fragments of the “global” dataset that represents the material record, or even the record of any specific time period or region. If one takes any definition of “Big Data” as it is generally understood, a corpus of information which is too massive for desktop-based or manual analysis or manipulation, no single archaeological dataset is likely to have these attributes of size and scale. The significance of Big Data for archaeology lies not so much in the analysis and manipulation of single or multiple collections of vast datasets but rather in the bringing together of multiple data, created at different times, for different purposes and according to different standards; the interpretive and critical frameworks needed to create knowledge from them. Archaeology is “Big Data” in the sense that it is “data that is bigger than the sum of its parts.”

Those parts are massively varied. Data in archaeology can be normal photographic images, images and data from remote sensing, tabular data of information such as artifact findspots, numerical databases, or text. It should also be noted that the act of generating archaeological data is rarely, if ever, the end of the investigation or project. Any dataset produced in the field or the lab typically forms part of a larger interpretation and interpolation process and – crucially – archaeological data is often not published in a consistent or interoperable manner; although approaches to so-called Grey Literature, which constitutes reports from archaeological surveys and excavations that typically do not achieve a wide readership, are discussed below. This fits with a general characteristic of Big Data, as opposed to the “e-Science/ Grid Computing” paradigm of the 2000s. Whereas the latter was primarily concerned with “big infrastructure,” anticipating the need for

scientists to deal with a “deluge” of monolithic data emerging from massive projects such as the Large Hadron Collider, as described by Tony Hey and Anne Trefethen, Big Data is concerned with the mass of information which grows organically as the result of the ubiquity of computing in everyday life and in everyday science. In the case of archaeology, it may be considered more as a “complexity deluge,” where small data, produced on a daily basis, forms part of a bigger picture.

There are exceptions: Some individual projects in archaeology are concerned with terabyte-scale data. The most obvious example in the UK is the North Sea Paleolandscapes, led by the University of Birmingham, a project which has reconstructed the Early Holocene landscape of the bed of the North Sea, which was an inhabitable landscape until its inundation between 20,000 and 8,000 BP – so-called Doggerland. Vince Gaffney and others describe drawing on 3D seismic data gathered during the process of oil prospection, this project has used large-scale data analytics and visualization to reconstruct the topography of the preinundation land surface spanning an area larger than the Netherlands, and to thus allow inferences as to what environmental factors might have shaped human habitation of it; although it must be stressed that there is no direct evidence at all of that human occupation. While such projects demonstrate the potential of Big Data technologies for conducting large-scale archaeological research, they remain the exception. Most applications in archaeology remain relatively small scale, at least in terms of the volume of data that is produced, stored, and preserved.

However, this is not to say that approaches which are characteristic of Big Data are not changing the picture significantly in archaeology, especially in the field of landscape studies. Data from geophysics, the science of scanning subterranean features using techniques such as magnetometry and resistivity typically produce relatively large datasets, which require holistic analysis in order to be understood and interpreted. This trend is accentuated by the rise of more sophisticated data capture techniques in the field, which is increasing the capacity of data that can be gathered and analyzed. Although still not “big” in the literal sense

of “Big Data,” this class of material undoubtedly requires the kinds of approaches in thinking and interpretation familiar from elsewhere in the Big data agenda. Recent applications in landscape archaeology have highlighted the need both for large capacity and interoperation. For example, integration of data from the in the Stonehenge Hidden Landscape project, also directed by Gaffney, provides for “seamless” capture of reams of geophysical data from remote sensing, visualizing the Neolithic landscape beneath modern Wiltshire to a degree of clarity and comprehensiveness that would only have been possible hitherto with expensive and laborious manual survey. Due to improved capture techniques, this project succeeded in gathering a quantity of data in its first two weeks equivalent to that of the landmark Wroxeter survey project in the 1990s.

These early achievements of big data in an archaeological context fall against a background of falling hardware costs, lower barriers to usage, and the availability of generic web-based platforms where large-scale distributed research can be conducted. This combination of affordability and usability is bringing about a revolution in applications such as those described above, where remote sensing is reaching new concepts and applications. For example, coverage of freely available satellite imagery is now near-total; graphical resolution is finer for most areas than ever before (1 m or less); and pre-georeferenced satellite and aerial images are delivered to the user’s desktop, removing the costly and highly specialized process of locating imagery of the Earth’s surface. Such platforms also allow access to imagery of archaeological sites in regions which are practically very difficult or impossible to survey, such as Afghanistan, where declassified CORONA spy satellite data are now being employed to construct inventories of the region’s (highly vulnerable) archaeology. If these developments cannot be said to have removed the boundaries within which archaeologists can produce, access, and analyze data, then it has certainly made them more porous.

As in other domains, strategies for the storage and preservation of data in archaeology have a fundamental relationship with relevant aspects of

the Big Data paradigm. Much archaeological information lives on the local servers of institutions, individuals, and projects; this has always constituted an obvious barrier to their integration into a larger whole. However, weighing against this is the ethical and professional obligation to share, especially in a discipline where the process of gathering the data (excavation) destroys its material context. National strategies and bodies encourage the discharge of this obligation. In the UK, as well as data standards and collections held by English Heritage, the main repository for archaeological data is the Archaeology Data Service, based at the University of York. The ADS considers for accession any archaeological data produced in the UK in a variety of formats. This includes most of the data formats used in day-to-day archaeological workflows: Geographic Information System (GIS) databases and shapefiles, images, numerical data, and text. In the latter case, particular note should be given to the “Grey Literature” library of archaeological reports from surveys and excavations, which typically present archaeological information and data in a format suitable for rapid publication, rather than the linking and interoperation of that data. Currently, the Library contains over 27,000 such reports. Currently, the total volume of the ADS’s collections stands at 4.5 Tb (I thank Michael Charno for this information). While this could be considered “big” in terms of any collection of data in the humanities, it is not of a scale which would overwhelm most analysis platforms; however what is key here is that it is most unlikely to be useful to perform any “global” scale analysis across the entire collection. The individual datasets therein relate to each other only inasmuch as they are “archaeological.” In the majority of cases, there is only fragmentary overlap in terms of content, topic, and potential use. A 2007 ADS/English Heritage report on the challenges of Big Data in archaeology identified four types of data format potentially relevant to Big Data in the field: LIDAR (Light Detection and Ranging or Laser Imaging Detection and Ranging) data, which models terrain elevation modelled from airborne sensors, 3D laser scanning, maritime survey, and digital video. At first glance this appears

to underpin an assumption that the primary focus is data formats which convey larger individual data objects, such as images and geophysics data, with the report noting that “many formats have the potential to be Big Data, for example, a digital image library could easily be gigabytes in size. Whilst many of the conclusions reached here would apply equally to such resources this study is particularly concerned with Big Data formats in use with technologies such as lidar surveys, laser scanning and maritime surveys.”

However, the report also acknowledges that “If long term preservation and reuse are implicit goals data creators need to establish that the software to be used or toolsets exist to support format migration where necessary.” It is true that any “Big Data” which is created from an aggregation of “small data” must interoperate. In the case of “social data” from mobile devices, for example, location is a common and standardizable attribute that can be used to aggregate Tb-scale datasets: heat maps of mobile device usage can be created which show concentrations of particular kinds of activity in particular places at particular times. In more specific contexts hashtags can be used to model trends and exchanges between large groups. Similarly intuitive attributes that can be used for interoperation, however, elude archaeological data, although there is much emerging interest in Linked Data technologies, which allow the creation of linkages between web-exposed databases, provided they conform (or can be configured to conform) to predefined specifications in descriptive languages such as RDF. Such applications have proved immensely successful in areas of archaeology concerned with particular data types, such as geodata, where there is a consistent base reference (such as latitude and longitude). However, this raises a question which is fundamental to archaeological data in any sense. Big Data approaches here, even if the data is not “Big” in terms of relative terms to the social and natural sciences, potentially allows an “n=all” picture of the data record. As noted above, however, this record represents only a tiny fragment of the entire picture. A key question, therefore, is does “Big data” thinking risk technological determination, constraining what

questions can be asked? This is a point which has concerned archaeologists since the very earliest days of computing in the discipline. In 1975, a skeptical Sir Moses Finley noted that “It would be a bold archaeologist who believed he could anticipate the questions another archaeologist or a historian might ask a decade or a generation later, as the result of new interests or new results from older researchers. Computing experience has produced examples enough of the unfortunate consequences . . . of insufficient anticipation of the possibilities at the coding stage.”

Conclusion

Such questions probably cannot be predicted, but big data is (also) not about predicting questions. The kind of critical framework that Big Data is advancing, in response to the ever-more linkable mass of pockets of information, each themselves becoming larger in size as hardware and software barriers lower, allows us to go beyond what is available “just” from excavation and survey. We can look at the whole landscape in greater detail and at new levels of complexity. We can harvest public discourse about cultural heritage in social media and elsewhere and ask what that tells us about that heritage’s place in the contemporary world. We can examine what are the fundamental building blocks of our knowledge about the past and ask what do we gain, as well as lose, by putting them into a form that the World Wide Web can read.

References

- Archaeology data service. <http://archaeologydataservice.ac.uk>. Accessed 25 May 2017.
- Austin, T., & Mitcham, J. (2007). *Preservation and management strategies for exceptionally large data formats: ‘Big Data’*. Archaeology Data Service & English Heritage: York, 28 Sept 2007.
- Gaffney, V., Thompson, K., & Finch, S. (2007). *Mapping Doggerland: The Mesolithic landscapes of the Southern North Sea*. Oxford: Archaeopress.
- Gaffney, C., Gaffney, V., Neubauer, W., Baldwin, E., Chapman, H., Garwood, P., Moulden, H., Sparrow, T., Bates, R., Löcker, K., Hinterleitner, A., Trinks, I., Nau, W., Zitz, T., Floery, S., Verhoeven, G., & Doneus, M. (2012). The Stonehenge Hidden Landscapes Project. *Archaeological Prospection*, 19(2), 147–155.
- Tudhope, D., Binding, C., Jeffrey, S., May, K., & Vlachidis, A. (2011). A STELLAR role for knowledge organization systems in digital archaeology. *Bulletin of the American Society for Information Science and Technology*, 37(4), 15–18.

Artificial Intelligence

Feras A. Batarseh
College of Science, George Mason University,
Fairfax, VA, USA

Synonyms

AI; Intelligent agents; Machine intelligence

Definition

Artificial Intelligence (often referred to as AI) is a field in computer science that is concerned with the automation of intelligence and the enablement of machines to achieve complex tasks in complex environments. This definition is an augmentation of two preexisting commonplace AI definitions (Goebel et al. 2016; Luger 2005).

AI is an umbrella that has many subdisciplines, big data analytics is one of them. The traditional promise of machine intelligence is being partially rekindled into a new business intelligence promise through big data analytics.

This entry covers AI, and its multiple subdisciplines.

Introduction

AI is a field that is built on centuries of thought; however, it became a recognized field for only over 70 years or so. AI is challenged in many ways, identifying what’s *artificial* versus what is *real* can be tricky in some cases, for example: “A tsunami is a large wave in an ocean caused by an

earthquake or a landslide. Natural tsunamis occur from time to time. You could imagine an artificial tsunami that was made by humans, by exploding a bomb in the ocean for instance, yet, it still qualifies as a tsunami. One could also imagine fake tsunamis: “using computer graphics, or natural, for example, a mirage that looks like a tsunami but is not one.” (Poole and Mackworth 2010).

However, intelligence is arguably different: you cannot create an illusion of intelligence or fake it. When a machine acts intelligently, it is then *intelligent*. There is no known way that a machine would demonstrate intelligence randomly.

The field of AI continuously poses a series of questions: How to define or observe intelligence? Is AI safe? Can machines achieve super-intelligence? among many other questions. In his famous manuscript, “Computing Machinery and Intelligence” (Turing 1950), Turing paved the way for many scientists to think about AI through answering the following: *Can Machines think?* To be able to imitate, replicate, or augment human intelligence, it is crucial to first understand what intelligence exactly means. For that, AI becomes a field that overlaps other areas of study, such as biology (the ability to understand the human brain and nervous system); philosophy is another field that has been highly concerned with AI (understanding how AI would affect the future of humanity – among many other philosophical discussions).

AI Disciplines

There have been many efforts towards achieving intelligence in machines that has led to the creation of many disciplines in AI, such as:

1. Machine Learning: is when intelligent agent learn by exploring their surroundings and while figuring out what actions are the most rewarding.
2. Neural Networks: are a learning paradigm inspired by the human nervous system. In
- neural networks, information is processed by a set of interconnected nodes called neurons.
3. Genetic Algorithms (GA): is a method that finds a solution or an approximation to the solution for optimization and search problems. GAs use biological techniques such as mutation, crossover, and inheritance.
4. Natural Language Processing (NLP): is a discipline that deals with linguistic interactions between humans and computers. It is an approach dedicated to improving the human-computer interaction. This approach is usually used for audio recognition.
5. Knowledge-based systems (KBS): are intelligent systems that reflect the knowledge of a proficient person, also referred to as expert systems. KBS are known to be one of the earliest disciplines of modern AI.
6. Computer Vision: is a discipline that is concerned with injecting intelligence to enforce the ability of perceiving objects. It occurs when the computer captures and analyzes images of the 3D world. This includes making the computer recognize objects in real-time.
7. Robotics: is a central field of AI that deals with building machines that imitate human actions and reactions. Robots in some cases have human features such as arms and legs, and in many other cases, are far from how humans look like. Robots are referred to as intelligent agents in some instances.
8. Data Science and Advanced Analytics: is a discipline that aims to increase the level of data-driven decision-making and providing improved descriptive and predictive pointers. This entry has been the focus of recent business AI applications, to the degree that many interchangeably (wrongly though) refer to it as AI. Many organizations are adopting this area of research and development. It has been used in many domains (such as healthcare, government, and banking). Intelligence methods are applied to structured data, and results are usually

presented in what is referred to as a data visualization (using tools such as Tableau, R, SPSS, and PowerBI).

Computer agents are a type of intelligent system that can interact with humans in a realistic manner. They have been known to beat the world's best chess player and locate hostages in a military operation. A computer agent is an autonomous or semiautonomous entity that can emulate a human. It can be either physical such as a robot or virtual such as an avatar. The ability to learn should be part of any system that claims intelligence.

AI Challenges and Successes

Intelligent agents must be able to adapt to changes in their environment. Such agents, however, have been challenged by many critics and thinkers for many reasons. Major technical and philosophical challenges to AI include: (1) The devaluation of humans: many argue that AI would replace humans in many areas (such as jobs and day-to-day services). (2) The lack of hardware that can support AI's extensive computations. Although Moore's law sounds intriguing (which states that the number of registers on an integrated circuit is doubling every year), that is still a fairly slow pace for what AI is expected to require in terms of hardware. (3) The effect of AI: Whenever any improvement in AI is accomplished, it is disregarded as a calculation in a computer that is driven by a set of instructions, and not real intelligence. This was one of the reasons the AI winter occurred (lack of research funding in the field). The field kept providing exploratory studies but there was a lack of real applications to justify the funding. Recently however, with technologies such as Deep Blue and Watson, AI is gaining attention and attracting funding (in academia, government, and the industry). (4) Answering the basic questions of when and how to achieve AI. Some researchers are looking to achieve Artificial General Intelligence (AGI) or

Superintelligence, which is a form of intelligence that can continuously learn and replicate human's thought, understand context, develop emotions, intuitions, fears, hopes, and reasoning skills. That is a much wider goal of AI than the existing Narrow-intelligence, which presents machines that have the ability perform a predefined set of tasks intelligently. Narrow intelligence is currently being deployed in many applications such as driverless cars and intelligent personal assistants. (5) Turing's list of potential AI roadblocks: presented in his famous paper, those challenges are still deemed relevant (among many other potential challenges).

In spite of the listed major five challenges, AI already presented multiple advantages such as: (1) greater calculation precision, accuracy, and the lack of errors, (2) performing tasks that humans are not able to or ones that are deemed too dangerous (such as space missions, and military operations), (3) accomplishing very complex tasks such as fraud detection, events prediction, and forecasting. Furthermore, AI had many successful deployments such as: Deep Blue (a chess computer), Autonomous Cars (produced by Tesla, Google and other technology, and automotive companies), IBM's Watson (a jeopardy computer), and Intelligent Personal Assistants (such as Apple's Siri and Amazon's Alexa).

Conclusions

AI is a continuously evolving field; it overlaps with multiple other areas of research such as computer science, psychology, math, biology, philosophy, and linguistics. AI is both feared by many due to the challenges listed in this entry and loved by many as well due to its many technological advantages in critical areas of human interest. AI is often referred to as the next big thing, similar to the industrial revolution and the digital age. Regardless of its pros, cons, downfalls, or potential greatness, it is an interesting field that is worth exploring and expanding.

Further Reading

- Goebel, R., Tanaka, Y., & Wolfgang, W. (2016). Lecture notes in artificial intelligence series. In: *Proceedings of the ninth conference on artificial general intelligence*, New York.
- Luger, G. (2005). *Artificial intelligence, structures and strategies for complex problem solving* (5th ed.). Addison Wesley, ISBN: 0-321-26318-9.
- Poole, D., & Mackworth, A. (2010). *Artificial intelligence: Foundation of computer agents* (1st ed.). Cambridge University Press, ISBN: 978-0-511-72946-1.
- Turing, A. M. (1950). Computing machinery and intelligence. *Journal of the Mind*, 59, 433–460.

Arts

Marcienne Martin

Laboratoire ORACLE [Observatoire Réunionnais des Arts, des Civilisations et des Littératures dans leur Environnement] Université de la Réunion Saint-Denis France, Montpellier, France

Big Data is a procedure that allows anyone connected to the Internet to access data whose content is as varied as the perception that every human being can have of objects of the world. This is true for the art.

Art is a cognitive approach that is applied to the objects in the world, which is quite unique because it uses the notion of “qualia,” which is the qualitative aspect of a particular experience: “Qualia are such things as colors, places and times” (Dummett 1978). Moreover, in the myth of Platon’s cavern the concept of “beautiffulness” belongs to the world of Ideas. What is more, for Hegel, art is a sensitive representation of the truth approached through an individual form. In other words, art is a transcription of the objects of the Reality through the artistic sensibility of the author. This phenomenon is at the origin of new artistic currents giving direction for the writing of artwork, irrespective of the domain (painting, sculpture, writing, music . . .), as well as featuring performers who are in resonance with these new views about art. In painting, we mention Gothic art in connection with *Fra Angelico*, Renaissance

art with *Leonardo da Vinci*, *Michelangelo*, impressionism with *Claude Monet*, *Pierre Auguste Renoir*, *Edouard Manet*, and more recently cubism with *Georges Braque*, *Pablo Picasso*, *Lyonel Feininger*, *Fernand Leger*, and futurism with *Luigi Russolo*, *Umberto Boccioni*, just to name a few. There are also artists who have joined any artwork current as *Facteur Cheval* whose artistic work has been called a posteriori “naive art.”

Literary movements are part of a specific formatting of writing. This is the case with, for example, in France, Middle Ages with the epic or the courtly romance; in the nineteenth century, Romanticism with, in Germany, *the circle of Jena*; in England with *Lord Byron*’s works or else in the USA, *Edgar Allan Poe* who included the story of horror as artwork or *Herman Melville*’s novel with internationally known: *Moby Dick* (1851); in the twentieth century, various movements have emerged, including the new novel illustrated by *Alain Robbe-Grillet*’s works; in the United States, writers like *Scott Fitzgerald*, *Ernest Hemingway* belong to contemporary history. Science fiction is a new scriptural approach created from imagination with no relation to reality. In music, its writing is at the origin of the creation of innovative rhythms transcribed through diverse instruments. Monody and polyphony have created the song and the opera. Around the Classical Age (eighteenth century), the art of music is transcribed in the form of a sonata, symphony, string quartet or chamber music, etc. Popular music as jazz, rock and roll, etc., is an art form appreciated. From the bitonality to the polytonality, the art of music has been enhanced ad infinitum. Finally, architecture is an art form which values monuments and houses; it is the case with modern architecture founded by the French architect *Le Corbusier*.

Some theories in psychology consider art as an act that would allow the sublimation of unfortunate experiences. For example, *Frida Kahlo* transcribed her physical suffering in her paintings “The broken column” (1944). This technique, called “art therapy,” was developed based on the relationship between suffering and one’s ability to express oneself through art and, thus, sublimate

suffering, which expresses one's resilience capacity. For example, *Jean Dubuffet*, an artist, discovered people's artworks who were suffering from psychiatric disorders; he named this art form "Rough art." Other psychological theorists have analyzed this phenomenon.

Through its specific manifestations, art stands out from the usual human paradigms, like pragmatism (the objects of reality approached as such), representation (lexical-semantic fields, doxa, culture . . .), or their symbolization (flags, military decorations . . .). Art uses the fields of imagination, emotions, and the sensibility of the author, which makes each work unique in its essence.

If art is difficult to define in its specificity (quale or feeling), it can be analyzed in its manifestations with the informational decryption realized through various perspectives on the work by art critics, authors specialized in this area, magazines, etc. Goodman tried to find common invariants of a feeling to another concerning the same object, which he expressed as a matching criteria of the qualia by the following equation: $q(x) = q(y)$ ssi M_{xy} (q = quale, M = matching). Qualia are phenomena which belong to the domain of individual perception, which cannot be transmitted as such from one individual to another; this phenomenon refers to the concept of solipsism that covers the meaning of: "attitude of the thinking subject for which his own consciousness is the only reality, the other consciousnesses, the outside world are only as representations."

The rewriting of the concept of qualia through an informational system will consider the objects causing these feelings and not the qualia. The information system is the *substratum* from which the living world is based upon. Indeed, whatever the content of information is, its transfer from a transmitter X will influence the perception of the environment and the responses given to it for a receiver Y , which will result, sooner or later, in the "butterfly effect" discovered by Lorenz and developed by Gleick (1989). Furthermore, the oscillation of objects in the universe between entropy and negative entropy is articulated around the factor time; without it, neither the evolution of objects in the world would exist nor their

transformation; concerning the information coupled with the advancement of the living world, it would have no reality.

The exchange of information is a start that allows the world of the living to exist and perpetuate itself through time. The informational exchange has been the subject of numerous studies. The understanding of the nature of an object in the world is multifaceted: an unknown object generated as hypotheses and beliefs as postulates of more information than a known object. Norbert Wiener came up with a concept defined by the term "cybernetic"; this concept refers to the notion of the quantity of information which is connected to a classical notion in statistical mechanics, that is, entropy. As information in a system is a measure of the degree of organization, entropy is a measure of the degree of disruption of a system. One is simply the opposite of the other. In the world of common realities, the transfer of information is realized from an object X to an object Y , and this will affect particular environments, provided that such information be updated and sent to an object Z . In the virtual world, or the Internet, information is available to any user and it is consulted, generally, at the discretion of people; *a contrario*, in the living world, information is part of requirements that are linked to survival and continuity.

In addition, information on the Internet is reticular, which refers to the basic idea of points which communicate with each other, or differently with intersecting lines. The networked structure generates very different behavior from those in relation to social structure as tree-like or pyramidal type. As part of the "network of networks" or the Internet, each internet user occupies a dual role: the network node and the link. Indeed, from a point X (surfer), a new network can be created and can aggregate new members around a common membership (politics, art, fashion . . .) and mediated by tools called "social networks" such as Facebook and Twitter. The specificity of this type of fast growing networks can open on the discovery of an artist with his work put on the Internet as the South Korean artist *Psy* with his song "Gangnam Style" (December 2012) presented on *Youtube* (<https://www.youtube.com/watch?v=9bZkp7q19f0>).

The database on Internet has increased exponentially. At a new given information, feedback is given, and this ad infinitum. However, if the information is available, it is not solicited by each user of the Web. Only a personal choice directs the user to a particular type of information based on its own requests. The amount of existing information on the Internet as well as its storage, transfer, new feedback made after consultation by web users and their speed of transmission are part of the concept of “big data,” which is built around a double temporal link. Each user can connect to any Internet user regardless of where it is on the planet. The notion of time is in the immediacy. The amount of information available to each user is available to anyone, at any time, and regardless of the place of consultation, which refers to a timeless space.

To return to the concept of time, from the nineteenth century (era of industrialization) with the creation of transport rail or automobiles, distance perception by people has changed. The air transport is the source of a change in concepts of time and distance. Indeed, e.g., a Paris-New York travel is no longer expressed in the form of the distance between these two points, but in the length of time taken to reach them; thus Paris is 8 h from New York and not 7000 km. The tilting of the spatial dimension time in time dimension is in resonance with Internet where contact is part of immediacy: time and distance become one; they are redefined in the form of binary information through satellites and various computer media. The reticular structure of the digital society is composed of nodes and human links (Internet, experts in the field), but also of technological links and nodes (hardware, satellite, etc.).

Art is in resonance with the phenomenon of the rewriting of time and of space, with art called “ephemeral” in which the artist creates a work that will last only the time of its installation. The ephemeral art is a way of expressing time in the presence and focusing on the feeling, i.e., the quale, not the sustainability. This approach is the opposite of artworks whose purpose was to last beyond the generation that witnessed their creation. Examples would be Egyptian pyramids, the Venus of Milo, and the Mona Lisa. The Internet

is also the source of new approaches of art. This is the case of works by artists and which are retransformed by such another artist; two or more works can coexist while being retransformed in a third artwork. “In writing, painting, or what might be referred to as overlapped art, I could say that art is connected from my feeling to the creation of the other” (Martin 2014). *Mircea Bochis*, a Romanian artist, has created original videos mixing poetry of an author and a video created by another. After Dark is a new project for visiting a museum at night from his computer with the help of a robot. The National Gallery of British Art in London or Tate Britain has similar innovative projects.

Technological development has opened a new artistic approach to photography and filmography. Moreover, art has long been known in privileged social backgrounds and it was not until 1750 that the first French museum was opened to the public. If art in all its forms, through its digital rewriting (photographs, various montages, movies, videos, books online, etc.), is open to everyone, only personal choices will appeal to big data for consultation.

Further Reading

- Bochis, M. (2014). Artist. <http://www.bochis.ro/>. Accessed 15 August 2014.
- Botet Pradeilles, G., & Martin, M. (2014). *Apologie de la névrose suivi de En écho*. Paris: L'Harmattan.
- Buci-Glucksmann, C. (2003). *Esthétique de l'éphémère*. Paris: Éditions Galilée.
- Denizeau, G. (2011). *Palais idéal du facteur cheval: Le palais idéal, le tombeau, les écrits*. Paris: Nouvelles Éditions Scala.
- Dokic, J., & Égré, P. L'identité des qualia et le critère de Goodman. http://j.dokic.free.fr/philosophy/goodman_de1.pdf, <https://fr.scribd.com/document/144118802/L-identite-des-qualia-et-le-critere-de-Goodman-pdf>.
- Dummett, M. (1978). *Truth and other enigmas*. Cambridge, MA: Harvard University Press.
- Gleick, J. (1989). *La Théorie du Chaos*. Paris: Flammarion.
- Hegel, G. F. W. (1835). *Esthétique, tome premier*. Traduction française de Ch. Bénard. (posth.). http://www.uqac.quebec.ca/zone30/Classiques_des_sciences_sociales/index.html.
- Herrera, H. (2003). *Frida: biographie de Frida Kahlo*. Paris: Livre de poche.

- Martin, M. (2017). *La nomination dans l'art – Étude des œuvres de Mircea Bochis, peintre et sculpteur*. Paris: Éditions L'Harmattan.
- Melville, H. (2011). *Moby Dick*. Paris: Editions Phébus.
- Platon. (1879). *L'État ou la République de Platon. Traduction nouvelle par Bastien, Augustin*. Paris: Garnier frères. <http://catalogue.bnf.fr/ark:/12148/cb31121998c>.
- Wiat, C. (1967). *Expression picturale et psychopathologie. Essai d'analyse et d'automatique documentaires (principe – méthodes – codification)*. Paris: Editions Doin.
- Wiener, N. (1948). Cybernetics or control and communication. In *The animal and the machine*. Cambridge, MA: MIT Press.

Asian Americans Advancing Justice

Francis Dalisay
Communication & Fine Arts, College of Liberal Arts & Social Sciences, University of Guam, Mangilao, GU, USA

Asian Americans Advancing Justice (AAAJ) is a national nonprofit organization founded in 1991. It was established to empower Asian Americans, Pacific Islanders, and other underserved groups, ensuring a fair and equitable society for all. The organization's mission is to promote justice, unify local and national constituents, and empower communities. To this end, AAAJ dedicates itself to develop public policy, educate the public, litigate, and facilitate in the development of grassroots organizations. Some of their recent accomplishments have included increasing Asian Americans and Pacific Islanders' voter turnout and access to polls, enhancing immigrants' access to education and employment opportunities, and advocating for greater protections of rights as they relate to the use of "big data."

The Civil Rights Principles for the Era of Big Data

In 2014, AAAJ joined a diverse coalition comprising of civil, human, and media rights groups,

such as the ACLU, the NAACP, and the Center for Media Justice, to propose, sign, and release the "Civil Rights Principles for the Era of Big Data." The coalition acknowledged that progress and advances in technology would foster improvements in the quality of life of citizens and help mitigate discrimination and inequality. However, because various types of "big data" tools and technologies – namely, digital surveillance, predictive analytics, and automated decision-making – could potentially ease the level in which businesses and governments are able to encroach upon the private lives of citizens, the coalition found it critical that such tools and technologies are developed and employed with the intention of respecting equal opportunity and equal justice.

According to civilrights.org (2014), the Civil Rights Principles for the Era of Big Data proposes five key principles: (1) stop high-tech profiling, (2) guarantee fairness in automated decisions, (3) maintain constitutional protections, (4) enhance citizens' control of their personal information, and (5) protect citizens from inaccurate data. These principles were intended to inform law enforcement, companies, and policy-makers about the impact of big data practices on racial justice and the civil and human rights of citizens.

1. Stop high-tech profiling. New and emerging surveillance technologies and techniques have made it possible to piece together comprehensive details on any citizen or group, resulting in an increased risk of profiling and discrimination. For instance, it was alleged that police in New York had used license plate readers to document vehicles that were visiting certain mosques; this allowed the police to track where the vehicles were traveling. The accessibility and convenience of this technology meant that this type of surveillance could happen without policy constraints. The principle of stopping high-tech profiling was thus intended to limit such acts through setting clear limits and establishing auditing procedures for surveillance technologies and techniques.

2. Ensure fairness in automated decisions. Today, computers are responsible for making critical decisions that have the potential to affect the lives of citizens' in the areas of health, employment, education, insurance, and lending. For example, major auto insurers are able to use monitoring devices to track drivers' habits, and as a result, insurers could potentially deny the best coverage rates to those who often drive when and where accidents are more likely to occur. The principle of ensuring fairness in automated decisions advocates that computer systems should be operating fairly in situations and circumstances such as the one described. The coalition had recommended, for instance, that independent reviews be employed to assure that systems are working fairly.
3. Preserve constitutional protections. This principle advocates that government databases must be prohibited from undermining core legal protections, including those concerning citizens' privacy and their freedom of association. Indeed, it has been argued that data from warrantless surveillance conducted by the National Security Agency have been used by federal agencies, including the DEA and the IRS, even though such data were gathered outside the policies that rule those agencies. Individuals with access to government databases could also potentially use them for improper purposes. The principle of preserving constitutional protections is thus intended to limit such instances from occurring.
4. Enhance citizens' control of their personal information. According to this principle, citizens should have direct control over how corporations gather data from them, and how corporations use and share such data. Indeed, personal and private information known and accessible to a corporation can be shared with companies and the government. For example, unscrupulous companies can find vulnerable customers through accessing and using highly targeted marketing lists, such as one that might contain the names and contact information of citizens who have cancer. In this case, the principle of enhancing citizens' control of personal information ensures that the government and companies should not be able to disclose private information without a legal process to do so.
5. Protect citizens from inaccurate data. This principle advocates that when it comes to making important decisions about citizens – particularly, the disadvantaged (the poor, persons with disabilities, the LGBT community, seniors, and those who lack access to the Internet) – corporations and the government should work to ensure that their databases contain accurate of personal information about citizens. To ensure the accuracy of data, this could require disclosing the underlying data and granting citizens the right to correct information that is inaccurate. For instance, government employment verification systems have had higher error rates for legal immigrants and individuals with multiple surnames (including many Hispanics) than for other legal workers; this has created a barrier to employment. In addition, some individuals have lost job opportunities because of inaccuracies in their criminal history information, or because their information had been expunged.

The five principles above continue to help inspire subsequent movements highlighting the growing need to strengthen and protect civil rights in the face of technological change. Asian Americans Advancing Justice and the other members of the coalition also continue to advocate for these rights and protections.

Cross-References

- [American Civil Liberties Union](#)
- [Centers for Disease Control and Prevention \(CDC\)](#)

Further Reading

Civil rights and big data: Background material. <http://www.civilrights.org/press/2014/civil-rights-and-big-data.html>. Accessed 20 June 2016.

Association Analysis

► Data Mining

Association Versus Causation

Weiwu Zhang¹ and Matthew S. VanDyke²

¹College of Media and Communication, Texas Tech University, Lubbock, TX, USA

²Department of Communication, Appalachian State University, Boone, NC, USA

Scientific knowledge provides a general understanding of how the world is connected among one another. It is useful in providing a means of categorizing things (typology), a prediction of future events, an explanation of past events, and a sense of understanding about the causes of the phenomenon (causation). Association, also called correlation or covariation, is an empirical and statistical relationship between two variables such that changes in one variable are connected to changes in the other. However, association in and of itself does not necessarily imply a causal relationship between the two variables. It is only one of several necessary criteria for establishing causation. The other two criteria for causal relationships are time order and non-spurious relationships. While the advance of big data makes it possible and more effective to capture tremendous number of correlations and predictions than ever before, and statistical analyses may assess the degree of association between variables with continuous data analyzed from big datasets, one must consider the theoretical underpinning of the study and how data were collected (i.e., in a manner that measurement of an independent variable precedes measurement of a dependent variable) in order to determine if the causal relationship is valid.

The purpose of this entry is to focus on association and one function of scientific knowledge – causation, what they are, how they relate to and differ from each other, and how big data plays any role in this process.

Association

A scientific theory is the relationships between concepts or variables in ways that describe, predict, and explain how the world operates. One type of relationships between variables is association or covariation. In this relationship, changes in the values of one variable are related to changes in the values of the other variable. In other words, the two variables shift their values together. Some statistical procedures are needed to establish association. To determine whether variable A is associated with variable B, we must see how the values of variable B shift when two or more values of variable A occur. If values in variable B shift systematically with each of the levels of variable A, then we can say there is an association between variables A and B. For example, to determine whether aggressiveness is really associated with exposure to violent television programs, we must observe aggressiveness under at least two levels of exposure to violent television programs, such as high exposure and low exposure. If higher level of aggressiveness is found under the condition of higher exposure to violent television programs than under the condition of lower exposure, we can conclude a positive association between exposure to television violence and aggressiveness. If lower level of aggressiveness is observed under the condition of higher exposure to violent television programs than under the condition of lower exposure, we can conclude a negative or inverse association between the two variables. Both situations indicate that exposure to television violence and aggressiveness are associated or covary.

To claim that variable A is a cause of variable B, the two variables must be associated with one another. If high- and low viewing of violent programs on television are equally related to level of aggressiveness, then there is no association between watching television violence and aggressiveness. In other words, knowing a person's viewing of violent programs on television does not help in any way predicting a person's level of aggressiveness. In this case, watching television violence cannot be a cause of aggressiveness. On the other hand, simple association between these

two variables does not imply causation. Other criteria are needed to establish causation.

A dominant theoretical framework in media communication research is the agenda-setting theory. McCombs and colleagues' research suggests that there is an association between prominent media coverage and what people tend to think about. That is, media emphasis on certain issues tends to be associated with the perceived importance of issues among the public. Recent research has examined the agenda-setting effect in the context of big data, for example, assessing the relationship between digital content produced by traditional media outlets (e.g., print; television) and user-generated content (i.e., blogs, forums, and social media). While agenda-setting research typically identifies associations between the prominence of media coverage of some issues and the importance public attaches to those issues, research designs must account for the sequence (i.e., time order) in which variables occur. For example, while it is plausible to think that media coverage influences what the public thinks about, in the age of new media, the public also plays increasingly important role in influencing what is covered by the news media outlets. Such explorations are questions of causality and would require a consideration of time order sequence between variables. Additionally, potential external causes of variation must be considered in order to truly establish causation.

Time Order

A second criterion for establishing causality is that a cause (independent variable) should take place before its effect (dependent variable). This means that changes in the independent variable should influence changes in the dependent variable, but not vice versa. This is also called the direction of influence (from independent variable to dependent variable). For some relationships in social research, the time order or direction of influence is clear. For instance, one's parents' education always occurs before their children's education. For others, the time order is not easy to determine. For example, while it is easy to find that viewing

television violence and aggressiveness are related, it is much harder to determine which variable causes the changes in the other. One plausible explanation is that the more one views television violence, the more one imitates the violent behavior on television and becomes more aggressive (per social learning theory). An equally plausible interpretation is that an aggressive person is usually attracted to violent television programs. Without any convincing evidence about the time order or direction of influence, there is no sound basis for determining which is the cause (independent variable) and which is the effect (dependent variable).

Some research designs such as controlled experiment are easier to decide on the time order of influence. Recent research examining people's use of mobile technology employed a field experiment to understand people's political web-browsing behavior. For example, Hoffman and Fang tracked individuals' web-browsing behavior over 4 months to determine predictors (e.g., political ideology) of the amount of time individuals spend browsing certain political content over others. Such research is able to establish that some pre-existing characteristic predicts or manipulation causes a change to the outcome of web-browsing behavior.

Non-spurious Relationships

This is the third essential criterion for establishing a causal relationship: when a relationship between two variables is not caused by variation in a third or extraneous variable. This means that the seeming association between two variables might be caused by a common third or extraneous variable (spurious relationship) rather than an influence of the presumed independent variable on the dependent variable. One well-known example is the association between a person's foot size and one's verbal ability in the 2010 US Census. If you believe that association or correlation implies causation, then you might think so. But the apparent relationship between one's foot size and verbal ability is a spurious one because one's foot size and verbal ability is linked to a common third

variable – age. As one grows older, one’s foot size becomes larger and as one grows older, one becomes better at communicating, but there is no logical and inherent relationship between foot size and verbal ability. To return to the agenda-setting example, perhaps a third variable would influence the relationship between media issue coverage and the importance public attaches to issues. For example, perhaps the nature of issue coverage (e.g., emotional coverage; coverage of issues of personal importance) would influence what the public thinks about issues presented by the media. Therefore, when we infer a causal relationship from an observed association, we need to rule out the influence of a third variable (or rival hypothesis) that might have created a spurious relationship between the variables.

In conclusion, despite the accumulation of enormous number of associations or correlations in the era of big data, association still does not supersede causation. To establish causation, the criteria of time order and non-spurious relationships must also be met with sound theoretical foundation in the broader context of big data.

Cross-References

- ▶ [Association Analysis](#)
- ▶ [Correlation Versus Causation](#)
- ▶ [Social Sciences](#)

Further Reading

- Babbie, E. (2007). *The practice of social research* (11th ed.). Belmont: Wadsworth.
- Hermida, A., Lewis, S., & Zamith, R. (2014). Sourcing the Arab spring: A case study of Andy Carvin’s sources on Twitter during the Tunisian and Egyptian revolutions. *Journal of Computer-Mediated Communication*, 19(3), 479.
- Hoffman, L., & Fang, H. (2014). Quantifying political behavior on mobile devices over time: A user evaluation study. *Journal of Information Technology & Politics*, 11(4), 435.
- Mahrt, M., & Scharrow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, 57(1), 20.
- McCombs, M. (2004). *Setting the agenda: The mass media and public opinion*. Cambridge, UK: Polity.

- Reynolds, P. (2007). *A primer in theory construction*. Boston: Pearson/Allyn & Bacon.
- Shoemaker, P., Tankard, J., & Lasorsa, D. (2004). *How to build social science theories*. Thousand Oaks: Sage.
- Singleton, R., & Straits, B. (2010). *Approaches to social research* (5th ed.). New York: Oxford University Press.

Astronomy

R. Elizabeth Griffin

Dominion Astrophysical Observatory, British Columbia, Canada

Definition

The term “Big Data” is severally defined and redefined by many in the fields of scientific observations and the curation and management thereof. Probably first coined in reference to the large volumes of images and similar information-rich records promised by present-day and near-future large-scale, robotic surveys of the night sky, the term has come to be used in reference to the data that result from almost any modern experiment in astronomy, and in so doing has lost most of the special attributes which it was originally intended to convey. There is no doubt that “big” is only relative, and scientific data have always presented the operator with challenges of size and volume, so a reasonable definition is also a relative one: “big data” refers to any set or series of data that are too large to be managed by existing methods and tools without a major rethink and redevelopment of technique and technology, be they hardware or software.

According to the above definition, “big data” have always featured in astronomy. The early observers were acutely aware of visible changes manifested by certain stars, so the attendant need to make comparisons between observations called for an ability to recover past details. Those details necessarily involved accurate meta-data such as object ID, date, time, and location of the observer, plus some identifier for the observer. The catalogues of observations that were therefore kept (hand-written at first) needed to refer to object

names that were also catalogued elsewhere, so a chain of ever bigger data thus sprang up and developed. Hand-written catalogues gave way to typed or printed ones, each a work of substantial size; the Henry Draper Catalogue of positions and spectral types of the 225,000 brightest stars occupied nine volumes of the Harvard Annals between 1918 and 1924, and established a precedent for compiling and blending catalogued information to collate usefully the most up-to-date information available.

The pattern has continued and has expanded as space missions have yielded observations at wavelengths or frequencies not attainable from the ground, or have reached objects that are far too faint for inclusion in the earlier catalogues. New discoveries therefore bear somewhat unglamorous numerical identifiers that reflect the mission or survey concerned (e.g., Comet PanSTARRS C/2012 K1, nova J18073024 + 4,551,325, or pulsar PSR J1741–2054). Even devising and maintaining a comprehensive and unique nomenclature system is a challenge in itself to the “big data” emanating from multiobject spectroscopy and multichannel sweeps of the sky which are now being finalized for operation on large telescopes.

Astronomy’s hierarchical development in nomenclature belie an ability to resolve at all adequately either long-standing or newly minted mysteries which their data present. The brighter the star, the older the scheme for naming it, but that does not imply that astronomers have been able to work systematically through solving the puzzles posed by the brighter ones and that only the new observations recorded in the era of “big data” still await attention. Far from it. One of the most puzzling challenges in stellar astronomy involves a star of visual magnitude 2.9 (so it is easily visible to the unaided eye); many stars of magnitudes 3, 4, or 5 also present problems of multiplicity, composition, evolution, or status that call for many more data at many different wavelengths before some of the unknowns can confidently be removed. “Big data” in that sense are therefore needed for all those sorts and conditions, and even then astronomers will probably find good reason to call for yet more.

The concept of “big data” came vividly to the fore at the start of the twenty-first century with the

floating of an idea to federate large sets of digital data, such as those produced by the Sloan Digital Survey, the Hubble Space Telescope, or 2MASS, in order to uncover new relationships and unexpected correlations which then-current thinking had not conceived. Even if the concept was not unique in science at the time, it could be soluble only in astronomy because of the highly proficient schemes for data management that then existed uniquely in that science (and which continue to lead other sciences). The outcome – the *Virtual Observatory* (VO) – constitutes an ideal whereby specified data sets, distributed and maintained at source, can be accessed remotely and filtered according to specified selection criteria; as the name implies, the sources of the observations are data sets rather than telescopes. But since the astronomical sources involved had resulted from multinational facilities, national VO projects soon became aligned under an International VO Alliance. The ideal of enabling data from quite disparate experiments to be merged effectively required adherence to certain initial parameters such as data format and descriptors, and the definition of minimum meta-data headings; those are now set out in the “VO Table.”

Because objects in the cosmos can change, on any time-scale, and either periodically or unexpectedly (or both), the matter of storing astronomy’s “big data” has posed prime storage challenges to whichever age bred the equipment responsible. Storage deficits have always been present, and simply assumed different forms depending on whether the limiting factors were scribes to prepare hand-written copies, photographic plates that were not supported by equipment to digitize efficiently all the information on them, or (in pre-Internet days) computers and magnetic tapes of sufficient capacity and speed to cope with the novel demands of CCD frames. Just as modern storage devices now dwarf the expectations of the past, so there is a comfortable assumption that expansions in efficiency, tools and technologies will somehow cope with ever increasing demands of the future (“Moore’s law”); certainly the current lack of proven adequate devices has not damped astronomy’s enthusiasm for multiscale surveys, nor required the planners to keep within data-storage bounds that can actually be *guaranteed* today.

An important other side to the “data deluge” coin is the inherent ability to share and reuse, perhaps in an interdisciplinary application, whatever data are collected. While astronomers themselves, however ambitious, may not have all the resources at their command to deal exhaustively on their own with all the petabytes of data which their experiments will deliver and will continue to deliver, the multifaceted burdens of interpretation can nowadays be shared by an ever broadening ensemble of brains, from students and colleagues in similar domains to citizen scientists by the thousand. Not all forms of data analysis can be tackled adequately or even correctly by all categories of willing analysts, though the *Galaxy Zoo* project has illustrated the very considerable potential of amateurs and nonastronomical manpower to undertake fundamental classification tasks at the expenditure of very modest training. Indeed, a potential ability to share data widely and intelligently is the chief saving grace in astronomy’s desire to observe more than it can conceivably manage.

Broad sharing of astronomical data has become a possibility and now a reality because ownership of the data has increasingly been defined as public, though that has not always been the case. Observing time has routinely been awarded to individual researchers or research groups on the basis of competitive applications, and (probably stemming from the era of photographic records, when an observation consisted of a tangible product that needed a home) such observations were traditionally regarded as the property of the observatory where they were recorded. Plate archivists kept track of loans to PIs while new observations were being analyzed, but returns were then firmly requested – with the result that, on the whole, astronomical plate stores are remarkably complete. Difficulties arose in the case of privately (as opposed to publicly or nationally) owned observatories, as rulings for publicly funded organizations were not necessarily accepted by privately funded ones, but those problems have tended to evaporate as many modern facilities (and all the larger ones) are multinationally owned and operated.

Authoritarianism

Layla Hashemi

Terrorism, Transnational Crime, and Corruption Center, George Mason University, Fairfax, VA, USA

Individuals in authoritarian societies typically lack freedom of assembly and freedom of the press, but the internet and social media have provided an important voice to many members of authoritarian societies. Social media allows individuals to connect with people of similar minds, share opinions, and find a powerful way to counter the isolation often associated with life in authoritarian societies. However, advances in data collection, sharing, and storage have also drastically reshaped the policies and practices of authoritarian regimes. Digital technologies have not only expanded opportunities for public expression and discussion, they have also improved government capabilities to surveil and censor users and content. This is a complex situation in which, in authoritarian and repressive contexts, technology can be used to stifle and silence dissenting voices. The development of facial recognition and surveillance technology, which have been used to counter crime, can also pose threats to privacy and freedom.

Illustrative of this are the minority populations of Uighurs in China who are strategically tracked and denied human rights and just treatment by the authoritarian government. Artificial intelligence and digital technology are key in state surveillance of the millions of members of this Chinese minority. One of the great challenges of addressing the use of digital technology and big data in authoritarian contexts is determining what kinds of data are available. These technologies often range from relatively simple information communication technologies (ICTs) such as short message service (SMS) or email to complex data and tools such as geolocation, machine learning, and artificial intelligence.

Residents of authoritarian societies have embraced different forms of social media for

personal and public expression. Currently, among the most popular is Twitter, presently used, for example, by nearly three million in Iran – despite the platform being banned in the country – as well as in other authoritarian countries such as Turkey (over 13 million users) and Saudi Arabia (over 12 million users). Examining the Twitter activity of millions of activists allows social movement researchers to conduct detailed analyses and determine sources of discontent and the events that trigger mobilization. Also, popular online campaigns (e.g., #MeToo and #BlackLivesMatter) have facilitated mobilization for social justice and public discussion of controversial topics such as sexual harassment, police brutality, and violence across borders. ICT was widely used during the Green Movement and the Arab Spring in the Middle East and North Africa, and media has often served as a means to speak truth to power and express public discontent. Before the establishment of digital technology, forms of media such as newspapers, radio, photography, and film were used by those living in authoritarian contexts to express discontent and grievances.

In the era of the internet, massive amounts of information and data can be shared rapidly at a global scale with few resources. The shift from publishers and legacy media to online platforms and internet communications has expanded the responsibilities of technology and communications regulation in nation-states and in international institutions and transnational corporations, emphasizing the need for increased corporate social responsibility. Even in authoritarian societies, the shift of power to regulate information beyond state actors demonstrates the growing role of technology and complex data flows in personal, corporate, and civic spheres in the digital era.

Further Reading

- Jumet, K. D. (2018). *Contesting the repressive state: Why ordinary Egyptians protested during the Arab spring*. New York: Oxford University Press.
- Kabanov, Y., & Karyagin, M. (2018). Data-driven authoritarianism: Non-democracies and big data. In D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, & O. Koltsova (Eds.), *Digital*

transformation and global society (Communications in computer and information science) (Vol. 858, pp. 144–155). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02843-5_12.

- Mechkova, Valeriya, Daniel Pemstein, Brigitte Seim, Steven Wilson. 2020. Digital Society Project Dataset v2.
- Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. New Haven/London: Yale University Press.

Authorship Analysis and Attribution

Patrick Juola

Department of Mathematics and Computer Science, McNulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Synonyms

[Authorship profiling](#); [Authorship verification](#); [Stylistics](#); [Stylometry](#)

Introduction

Authorship attribution is a text classification technique used to infer the authorship of a document. By identifying features of writing style in a document and comparing it to features from other documents, a human analyst or a computer can make a determination of stylistic similarity and thus of the plausibility of authorship by any specific person. There are many applications, including education (plagiarism detection), forensic science (identifying the author of a piece of evidence such as a threatening letter), history (resolving questions of disputed works), and journalism (identifying the true authors behind pen names), among others.

Theory of Authorship Attribution

Human language is a complex system that is underconstrained, in the sense that there are

normally many ways to express roughly the same idea. Writers and speakers are therefore forced to make (consciously or unconsciously) choices about the best way to express themselves in any given situation. Some of the choices are obvious – for example, what speakers of American Standard English call “chips” speakers of Commonwealth dialects call “crisps” (and their “chips” Americans call “French fries”). Some authors may use the passive voice often, while others largely avoid it. Sometimes the choice is less noticeable – when you set the table, do you set the fork “to” the left of the plate, “on” the left of the plate, or “at” the left of the plate? While all are grammatically (and semantically) correct, some people have a marked preference for one form over another. If this preference can be detected, it can provide evidence for or against authorship of another document that does or does not match this pattern of preposition use.

Examples of Authorship Attribution in Practice

After some proposals dating back to the nineteenth century [see Juola (2008) for some history], one of the first examples of authorship was the analysis by Mosteller and Wallace (1963) of *The Federalist Papers* and their authorship. They found, for example, that Alexander Hamilton never used the word “while” (he used the word “whilst” instead) while James Madison was opposite. More subtly, they showed that, though both men used the word “by,” Madison used it much more frequently. From word-based observations like this, Mosteller and Wallace were able to apply Bayesian statistics to infer the authorship of each of the anonymously published *The Federalist Papers*.

Binongo (2003) used a slightly different method to address the authorship of the 15th book in the *Oz* series. Originally by L. Frank Baum, the series was taken over after Baum’s death by another writer named Ruth Plumly Thompson. The 15th book, *The Royal Book of Oz*, was published during this gap and has been variously attributed to both authors. Binongo

analyzed the fifty most common words in the *Oz* novels as a whole (a collection of fairly simple words like “the,” “of,” “after,” “with,” “that,” and so forth) and was able to show via standard statistical techniques that Baum and Thompson had notably different writing styles and that *The Royal Book* clearly matched Thompson’s.

Among the highest profile authorship attribution analyses is Juola’s analysis of *The Cuckoo’s Calling*. Published under the pen name “Robert Galbraith,” an anonymous tip on Twitter suggested that the real author was J.K. Rowling of *Harry Potter* fame. Juola analyzed several different aspects of “Galbraith’s” writing style and showed that Galbraith had a very similar grammar to Rowling, used many of the same types of words as Rowling, had about the same complexity of vocabulary as Rowling, used morphemes in the same way Rowling did, and even put words together into pairs like Rowling did. The obvious conclusion, which he drew, is that Rowling was, in fact, Galbraith, a conclusion that Rowling herself confirmed a few days later.

How Does It work?

The general method in these cases (and many others) is very similar and relies on a well-known data classification framework. From a set of known documents, extract a set of features (e.g., Binongo’s features were simply the fifty most common words, while Juola’s features included the lengths of words, the set of all adjacent word pairs, and the set of letter clusters like the “tion” at the end of “attention”). These sets of features can then be used as elements to classify unknown works using a standard classification system such as a support vector machine, nearest-neighbor classifier, a deep learning system, or many others. Similarly, scholars have documented more than 1000 proposed feature types that could be used in such a system. Despite or perhaps because of the open-ended nature of this framework, the search for best practices and most accurate attribution methods continues.

One key component of this search is the use of bake-off style competitive evaluations, where

researchers are presented with a common data set and invited to analyze it. Juola (2008) describes a 2004 competition in detail. Other examples of this kind of evaluation include the Plagiarism Action Network (PAN) workshops held annually from 2011 to this writing (as of 2018) and the 2017 Forensic Linguistics Dojo sponsored by the International Association of Forensic Linguists. Activities like these help scholars identify, evaluate, and develop promising methods.

Other Related Problems

Authorship attribution as defined above is actually only one of the many similar problems. Scholars traditionally divide attribution problems into two types; the open-class problem and closed-class problem. In the (easier) closed-class problem, the analyst is told to assume that the answer is one of the known possible authors – for example, it must be Baum or Thompson, but it can't be an unknown third party. In the open-class variation, “none of the above” is an acceptable answer; the unknown document might have been written by someone else. Open-class problems with only one known author are typically referred to as “authorship verification” problems, as the problem is more properly framed as “was the unknown document written by this (known) person, or wasn't it?”. Authorship verification is widely recognized as being more difficult than simple attribution among a closed-class group.

Sometimes, authorship scholars are asked to infer not the identity but the characteristics of the author of a document. For example, was this document written by a man or a woman? Where was the author from? What language did the author grow up speaking? These questions and many others like them have been studied as part of the “authorship profiling” problem. Authorship profiling can be addressed using largely the same methods, for example, by extracting features from a large

collection of writings by men and a large collection of writings by women and then seeing which collection's features better matches the unknown document.

Conclusions

Authorship attribution and related problems is an important area of research in data classification. Using techniques derived from big data research, individual and group attributes of language can be identified and used to identify authors by the attributes their writings share.

Cross-References

► [Bibliometrics/Scientometrics](#)

Further Reading

- Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16, 9.
- Juola, P. (2008). Authorship attribution. *Foundations and trends® in information retrieval*, 1(3), 233–334.
- Juola, P. (2015). The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(Suppl. 1), fqv040.
- Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology*, 60(1), 9–26.
- Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58, 275.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60, 538.

Authorship Profiling

► [Authorship Analysis and Attribution](#)

Authorship Verification

► [Authorship Analysis and Attribution](#)

Automated Modeling/ Decision Making

Murad A. Mithani

School of Business, Stevens Institute of
Technology, Hoboken, NJ, USA

Big data promises a significant change in the nature of information processing, and hence, decision making. The general reaction to this trend is that the access and availability of large amounts of data will improve the quality of individual and organizational decisions. However, there are also concerns that our expectations may not be entirely correct. Rather than simplifying decisions, big data may actually increase the difficulty of making effective choices. I synthesize the current state of research and explain how the fundamental implications of big data offer both a promise for improvement but also a challenge to our capacity for decision making.

Decision making pertains to the identification of the problem, understanding of the potential alternatives, and the evaluation of those alternatives to select the ones that optimally resolve the problem. While the promise of big data relates to all aspects of decision making, it more often affects the understanding, the evaluation, and the selection of alternatives. The resulting implications comprise of the dual decision model, higher granularity, objectivity, and transparency of decisions, and the bottom-up decision making in organizational contexts. I explain each of these implications in detail to illustrate the associated opportunities and challenges.

With data and information exceeding our capacity for storage, there is a need for decisions to be made on the fly. While this does not imply that all decisions have to be immediate, our

inability to store large amounts of data that is often generated continuously suggests that decisions pertaining to the use and storage of data, and therefore the boundaries of the eventual decision making context, need to be defined earlier in the process. With the parameters of the eventual decision becoming an apriori consideration, big data is likely to overcome the human tendency of procrastination. It imposes the discipline to recognize the desired information content early in the process. Whether this entails decision processes that prefer immediate conclusions or if the early choices are limited to the identification of critical information that will be used for later evaluation, the dual decision model with a preliminary decision far removed from the actual decision offers an opportunity to examine the available alternatives more comprehensively. It allows decision makers to have a greater understanding of the alignment between goals and alternatives. Compare this situation to the recruitment model for a human resource department that screens as well as finalizes prospective candidates in a single round of interviews, or separates the process into two stages where the potential candidates are first identified from the larger pool and they are then selected from the short-listed candidates in the second stage. The dual decision model not only facilitates greater insights, it also eliminates the fatigue that can seriously dampen the capacity for effective decisions. Yet this discipline comes at a cost. Goals, values, and biases that are part of the early phase of a project can leave a lasting imprint. Any realization later in the project that was not deliberately or accidentally situated in the earlier context becomes more difficult to incorporate into the decision. In the context of recruitment, if the skills desired of the selected candidate change after the first stage, it is unlikely that the short-listed pool will rank highly in that skill. The more unique is the requirement that emerges in the later stage, the greater is the likelihood that it will not be sufficiently fulfilled. This tradeoff suggests that an improvement in our understanding of the choices comes at the cost of limited maneuverability of an established decision context.

In addition to the benefits and costs of early decisions in the data generation cycle, big data allows access to information at a much more granular level than possible in the past. Behaviors, attitudes, and preferences can now be tracked in extensive detail, fairly continuously, and over longer periods of time. They can in turn be combined with other sources of data to develop a broader understanding of consumers, suppliers, employees, and competitors. Not only can we understand in much more depth the activities and processes that pertain to various social and economic landscapes, higher level of granularity makes decisions more informed and, as a result, more effective. Unfortunately, granularity also brings with it the potential of distraction. All data that pertains to a choice may not be necessary for the decision, and excessive understanding can overload our capacity to make inferences. Imagine the human skin which is continuously sensing and discarding thermal information generated from our interaction with the environment. What if we had to consciously respond to every signal detected by the skin? It is this loss of granularity that comes through the human mind responsive only to significant changes in temperature that saves us from being overwhelmed by data. Even though information granularity makes it possible to know what was previously impossible, information overload can lead us astray towards inappropriate choices, and at worse, it can incapacitate our ability to make effective decisions.

The third implication of big data is the potential for objectivity. When a planned and comprehensive examination of alternatives is combined with a deeper understanding of the data, the result is more accurate information. This makes it less likely for individuals to come up to an incorrect conclusion. This eliminates the personal biases that can prevail in the absence of sufficient information. Since traditional response to overcome the effect of personal bias is to rely on individuals with greater experience, big data predicts an elimination of the critical role of experience. In this vein, Andrew McAfee and Erik Brynjolfson (2012) find that regardless of the level of experience, firms that extensively rely on data for decision making are, on average, 6% more profitable than their peers. This suggests that as decisions become

increasingly imbued with an objective orientation, prior knowledge becomes a redundant element. This however does not eliminate the value of domain-level experts. Their role is expected to evolve into individuals who know what to look for (by asking the right questions) and where to look (by identifying the appropriate sources of data). Domain expertise and not just experience is the mantra to identify people who are likely to be the most valuable in this new information age. However, it needs to be acknowledged that this belief in objectivity is based on a critical assumption: individuals endowed with identical information that is sufficient and relevant to the context, reach identical conclusions. Yet anyone watching the same news story reported by different media outlets knows the fallacy of this assumption. The variations that arise when identical facts lead individuals to contrasting conclusions are a manifestation of the differences in the way humans work with information. Human cognitive machinery associates meanings to concepts based on personal history. As a result, even while being cognizant of our biases, the translation of information into conclusion can be unique to individuals. Moreover, this effect compounds with the increase in the amount of information that is being translated. While domain experts may help ensure consistency with the prevalent norms of translation, there is little reason to believe that all domain experts are generally in agreement. The consensus is possible in the domains of physical sciences where objective solutions, quantitative measurements, and conceptual boundaries leave little ambiguity. However, the larger domain of human experience is generally devoid of standardized interpretations. This may be one reason that a study by the Economist Intelligence Unit (2012) found a significantly higher proportion of data-driven organizations in the industrial sectors such as the natural resources, biotechnology, healthcare, and financial services. Lack of extensive reliance on data in the other industries is symptomatic of our limited ability for consensual interpretation in areas that challenge the positivistic approach.

The objective nature of big data produces two critical advantages for organizations. The first is transparency. A clear link between data, information, and decision implies the absence of personal

and organizational biases. Interested stakeholders can take a closer look at the data and the associated inferences to understand the basis of conclusions. Not only does this promise a greater buy-in from participants that are affected by those decisions, it develops a higher level of trust between decision makers and the relevant stakeholders, and it diminishes the need for external monitoring and governance. Thus, transparency favors the context in which human interaction becomes easier. It paves the way for richer exchange of information and ideas. This in turn facilitates the quality of future decisions. But due to its very nature, big data makes replications rather difficult. The time, energy, and other resources required to fully understand or reexamine the basis of choices makes transparency not an antecedent but a consequence of trust. Participants are more likely to believe in transparency if they already trust the decision makers, and those that are less receptive to the choices remain free to accuse the process as opaque. Regardless of the comprehensiveness of the disclosed details, transparency largely remains a symbolic expression of the participants' faith in the people managing the process.

A second advantage that arises from the objective nature of data is decentralization. Given that decisions made in the presence of big data are more objective and require lower monitoring, they are easier to delegate to people who are closer to the action. By relying on proximity and exposure as the basis of assignments, organizations can save time and costs by avoiding the repeated concentration and evaluation of information that often occurs at the various hierarchical levels as the information travels upwards. So unlike the flatter organizations of the current era which rely on the free flow of

information, lean organizations of the future may decrease the flow of information altogether, replacing it with data-driven, contextually rich, and objective findings. In fact, this is imminent since the dual decision model defines the boundaries of subsequent choices. Any attempt to disengage the later decision from the earlier one is likely to eliminate the advantages of granularity and objectivity. Flatter organizations of the future will delegate not because managers have greater faith in the lower cadres of the organization but because individuals at the lower levels are the ones that are likely to be best positioned to make timely decisions. As a result, big data is moving us towards a bottom-up model of organizational decisions where people at the interface between data and findings determine the strategic priorities within which higher-level executives can make their call. Compare this with the traditional top-down model of organizational decisions where strategic choices of the higher executives define the boundaries of actions for the lower-level staff. However, the bottom-up approach is also fraught with challenges. It minimizes the value of executive vision. The subjective process of environmental scanning allows senior executives to imbibe their valued preferences into organizational choices through selective attention to information. It enables organizations to do what would be uninformed and at times, highly irrational. Yet it sustains the spirit of beliefs that take the form of entrepreneurial action. By setting up a mechanism where facts and findings run supreme, organization of the future may constrain themselves to do only what is measurable. Extensive reliance on data can impair our capacity to imagine what lies beyond the horizon (Table 1).

Automated Modeling/Decision Making, Table 1 Opportunities and challenges for the decision implications of big data

	Big data implication	Opportunity	Challenge
1.	Dual decision model	Comprehensive examination of alternatives	Early choices can constrain later considerations
2.	Granularity	In-depth understanding	Critical information can be lost due to information overload
3.	Objectivity	Lack of dependence on experience	Inflates the effect of variations in translation
4.	Transparency	Free-flow of ideas	Difficult to validate
5.	Bottom-up decision making	Prompt decisions	Impairment of vision

In sum, the big data revolution promises a change in the way individuals and organizations make decisions. But it also brings with it a host of challenges. The opportunities and threats discussed in this article reflect different facets of the implications that are fundamental to this revolution. They include the dual decision model, granularity, objectivity, transparency, and the bottom-up approach to organizational decisions. The table above summarizes how the promise of big data is an opportunity as well as a challenge for the future of decision making.

Cross-References

- [Big Data Quality](#)
- [Data Governance](#)
- [Decision Theory](#)

Further Reading

- Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662–679.
- Economist Intelligence Unit. (2012). *The deciding factor: Big data & decision making*. New York, NY, USA: Capgemini/The Economist.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 61–67.

Aviation

Kenneth Button
Schar School of Policy and Government, George Mason University, Arlington, VA, USA

Introduction

In 2018 globally there were some 38.1 million commercial flights carrying about 4.1 billion passengers. Air freight is also now one of the major contributors to the global supply chain. The

world's 1770 dedicated cargo aircraft did 255 billion ton-kilometers in 2017 and along with the cargo carried in the belly holds scheduled passenger planes, combined to carry about 40% of world trade by value. Aviation is a large, complex, multifaceted, capital-intensive industry. It requires considerable amounts of data and their efficient analysis to function safely and economically.

Big data are widely used by commercial aviation and in a variety of different ways. It is important for weather predictions, maintenance planning, crew scheduling, fare setting, and so on. It plays a crucial role in the efficiency of air navigation service providers and in the economics of airline operations. It also is important for safety and security of passengers and cargo.

Weather

Flying is an intrinsically dangerous activity. It involves a continual fight against gravity and confronting other natural elements, especially weather. As enhanced technology has allowed longer and higher flights over more difficult terrains, the demands for more accurate weather forecasts have grown. The predictions made for weather now involve high-altitude wind directions and intensity and, with flights of 19 hours or more possible, take a much longer perspective than in the past. Fog or very low ceilings can prevent aircraft from landing, and taking off while turbulence and icing are also significant in-flight hazards. Thunderstorms are a problem because of severe turbulence and icing, due to the heavy precipitation, as well as hail, strong winds, and lightning, which can cause severe damage to an aircraft in flight. Locally, prior knowledge of shifts in wind intensity and direction allow airports to plan for changes in runway use permitting advantage to be taken of prevailing headwinds.

Traditionally, forecasting largely relied on data gathered from ground stations and weather balloons that recorded changes in barometric pressure, current weather conditions, and sky condition or cloud cover and manual calculations using simple models for forecasting. Distance between reporting sites, limited measuring techniques, lack of

adequate mechanical computational equipment, and inadequate models resulted in poor reliability. It was only in 1955 with the advent of computer simulation that numerical weather predictions became possible. Today with satellite data gathering, as well as strategically placed instruments to measure temperature, pressure, and other parameters on the surface of the planet as well as the atmosphere, massive amounts of data are available in real time. Manipulating the vast data sets and performing the complex calculations necessary to modern numerical weather prediction then require some of the most powerful supercomputers in the world (Anaman et al. 2017).

Weather data is increasingly being combined with other big data sets. The European Aviation Safety Agency's Data4Safety program collects and gathers all data that may support the management of safety risks at European level. This includes safety reports (or occurrences), telemetry data generated by an aircraft via their flight data recorders, and surveillance data from air traffic, as well as weather data. Similarly, near real-time weather data can be used not only to enhance safety and to avoid the discomforts of turbulence but also to reduce airline costs by adjusting routings to minimize fuel consumption – aviation fuel is about 17% of a commercial airline's costs.

Maintenance and Monitoring

Aircraft are complex machines. Correct maintenance and repair are important – the “airplane's health.” To this end there are regularly scheduled inspections and services. This has been supplemented more recently by in-service monitoring and recording of plane's performance in flight. The aerospace industry is now utilizing the enormous amount of data transmitted via sensors embedded on airplanes to preempt problems (Chen et al. 2016). Boeing, for example, analyzes two million conditions daily across 4000 aircraft as a part of its Airplane Health Management system. Pratt and Whitney have fitted about 5000+ sensors on its PW1000G engines for the Bombardier C Series and are generating about 10 GB of data per second. It provides engineers with millions of pieces of information that can be used to inspect and

respond to emerging problems immediately after landing rather than waiting for the next scheduled service. For example, if a part needs to be replaced, the system can send a message and the part is available on landing, and plane turnaround times are reduced.

From the business perspective, this allows more rapid rescheduling of hardware and crew should the problem-solving and remedy require significant down time. As in most cases, it is not the big data themselves that is important, but it is rather the ability to look at the data through machine learning, data mining, and so on that allows relevant information to be extracted isolated and analyzed. This helps airlines make better commercial decisions as well as improving safety.

Business Management

On the demand side, since the world began to deregulate air transportation in the late 1970s, airlines have been increasingly free to set fares and cargo rates, and to determine their routes served and schedules. To optimize these, and thus their revenues, airlines use big data sets regarding their consumers behavior, and this includes data on individuals' computer searches even when no seats are booked (Carrier and Fiig 2018). Many millions of pieces of data are collected daily regarding the type of ticket individuals buy, the frequency individuals fly, where to and in what class of seat, and their special needs, if any. They also obtain information on add-on purchases, such as additional baggage allowances and meals. Added to this there are data available from credit card, insurance, and car rental companies, hotels, and other sectors whose products are marketed with a flight.

This enables airlines to build up profiles of their customer bases and tailor service/fare packages to these. For example, easyJet uses an artificially intelligent algorithm that determines seat pricing automatically, depending on demand and to allows it to analyze historical data to predict demand patterns up to a year in advance. United Airlines use their “collect, detect, act” protocol to analyze over 150 variables in each customer profile with the objective of enhancing their yield management model. Delta

Air Lines, United and other airlines have apps that permit customers to track their bags on their smartphones.

Airlines sell tickets in “buckets”; each bucket has a set of features and a fare associated with it. The features may include the size and rack of seats, the refreshment offered, the entertainment provided, the order of boarding the plane, access to airport lounges, etc. The buckets are released at fares that tend to rise as the takeoff date approaches, although the lack of sales on any higher fare seats may produce a reversion to a lower-fare bucket. The aim is to maximize fare revenues. Leisure travelers tend to pay for their own trips, plan travel well in advance, often book in units (e.g., as a family), and, because the fare is important them, will take up the cheaper seats offered early. Business travelers with less advanced knowledge of their itineraries, and because employers pay the fare, generally book much later and take higher-fare seats. The various buckets are released partly based on real-time demand but also on projected later demand.

Added to this, the more expensive seats often allow for no-penalty cancellations at the last minute. To maximize revenues, airlines have to predict the probability of no-shows. Getting this wrong leads to overbookings and denied boarding compensation for ticketed passengers for whom there are no seats. Further, recently there has been considerable unbundling of the products being sold by airlines, and a seat is now often sold separately to other services. For example, with lower-priced seats, a passenger may buy seat selection, boarding priority, the use of a dedicated security check, a baggage allowance (including that carried on the plane), and the like separate from the airline ticket itself. Big data that provide insights into the actions of passengers on previous flights is a major input into estimating the number of no-shows and the supplementary services that different types of passengers tend to purchase.

Big data, and the associated growth in computation power, also facilitates more commercially efficient networking of scheduled airline services and, in particular, have contributed to

the development of hub-and-spoke structures (Button 2002). This involves the use of interconnecting services to channel passengers moving from **A** to **B** along a routing involving an intermediate, hub airport, **C**. By consolidating traffic at **C** originating not only from **A** but also from **D**, **E**, **F**, etc. and then taking them together to their final destination **B**, fewer total flights are needed and larger, more economically efficient planes may be used. To make the system efficient, there is a need for all the incoming flights to arrive at approximately the same time and to leave at approximately the same time – i.e., scheduling banks of flights. This is complex partly because of the needs to coordinate the ticketing of passengers interchanging at the hub and partly because there is a need to ensure cabin crews, with adequate flying hours, are free and planes are available for ongoing flights. The costs of missed connections are high, financially for the airline and in terms of wasted time for travelers. Weather forecasting and monitoring of aircraft performance have improved the reliability of hubbing as well as of passenger management.

In terms of cargo aviation, air transportation carries considerably amounts of high-value, low-volume consignments. This has been the result of freer markets and technical improvements in such things as the size and fuel economy of aircraft, the use of air cargo containerization, and a general shift to higher-value manufacture production. But at the forefront of it has been the development of computerized information systems that, just like passenger air transportation, allows easier and more knowledgeable booking, management of demand by differential pricing, tracking of consignment, and allocation of costs. Big data have been the key input to this dynamic logistics chain. It has allowed aviation to be a central player in the success of business using just-in-time production practices and has been a major factor in the growth of companies like FedEx (the largest cargo airline with 681 planes that did 17.5 billion freight ton-kilometers in 2018), UPS, and other express package carriers. The big aviation data is linked to that available for connection modes, such as trucking,

and for warehouse inventories to provide seamless supply chains.

interface of military and civilian aviation where their airspace needs overlap.

...and the Military

Inevitably the military side of aviation is also a large user of big data. Equally, and inevitably for security reasons, we know less about it. The objectives of the military are not commercial, much of the equipment used is often very different, and the organization is based upon command-and-control mechanisms rather than on the price mechanism (Hamilton and Kreuzer 2018). Nevertheless, and disregarding the ultimate motivations of military aviation, some aspects of its use of big data are similar to that in the civilian industry. Military aviation has similar demands for reliable weather forecasting and for monitoring the health of planes. It also uses it for managing its manpower and its air bases. And there is an inevitable

Further Reading

- Anaman, K. A., Quaya, R., & Owusu-Brown, B. (2017). Benefits of aviation weather services: A review of the literature. *Research in World Economy*, 8, 45–58.
- Button, K. J. (2002). Airline network economics. In D. Jenkins (Ed.), *Handbook of airline economics* (2nd ed., pp. 27–34). New York: Aviation Week.
- Carrier, E., & Fiig, T. (2018). Special issue: Future of airline revenue management. *Journal of Revenue and Pricing Management*, 17, 45–120.
- Chen, J., Lyu, Z., Liu, Y., Huang, J., Zhang, G., Wang, J., & Chen, X. (2016) A big data analysis and application platform for civil aircraft health management. In *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, Taipei.
- Hamilton, S. P., & Kreuzer, M. P. (2018). The big data imperative: Air force intelligence for the information. *Air and Space Power Journal*, 32, 4–18.

B

BD Hubs

- [Big Data Research and Development Initiative \(Federal, U.S.\)](#)

BD Spokes

- [Big Data Research and Development Initiative \(Federal, U.S.\)](#)

Behavioral Analytics

Lourdes S. Martinez
School of Communication, San Diego State
University, San Diego, CA, USA

Behavioral analytics can be conceptualized as a process involving the analysis of large datasets comprised of behavioral data in order to extract behavioral insights. This definition encompasses three goals of behavioral analytics intended to generate behavioral insights for the purposes of improving organizational performance and decision-making as well as increasing understanding of users. Coinciding with the rise of big data and the development of data mining techniques, a variety of fields stand to benefit from the emergence of behavioral analytics and its implications.

Although there exists some controversy regarding the use of behavioral analytics, it has much to offer organizations and businesses that are willing to explore its integration into their models.

Definition

The concept of behavioral analytics has been defined by Montibeller and Durbach as an analytical process of extracting behavioral insights from datasets containing behavioral data. This definition is derived from previous conceptualizations of the broader overarching idea of business analytics put forth by Davenport and Harris as well as Kohavi and colleagues. Business analytics in turn is a subarea within business intelligence and described by Negash and Gray as systems that integrate data processes with analytics tools to demonstrate insights relevant to business planners and decision-makers. According to Montibeller and Durbach, behavioral analytics differs from traditional descriptive analysis of behavioral data by focusing analyses on driving action and improving decision-making among individuals and organizations. The purpose of this process is threefold. First, behavioral analytics facilitates the detection of users' behavior, judgments, and choices. For example, a health website that tracks the click-through behavior, views, and downloads of its visitors may offer an opportunity to personalize user experience based on profiles of different types of visitors.

Second, behavioral analytics leverages findings from these behavioral patterns to inform decision-making at the organizational level and improve performance. If personalizing the visitor experience to a health website reveals a mismatch between certain users and the content provided on the website's navigation menu, the website may alter the items on its navigation menu to direct this group of users to relevant content in a more efficient manner. Lastly, behavioral analytics informs decision-making at the individual level by improving judgments and choices of users. A health website that is personalized to unique health characteristics and demographics of visitors may help users fulfill their informational needs so that they can apply the information to improve decisions they make about their health.

Applications

According to Kokel and colleagues, the largest behavioral databases can be found at Internet technology companies such as Google as well as online gaming communities. The sheer size of these datasets is giving rise to new methods, such as data visualization, for behavioral analytics. Fox and Hendler note the opportunity in implementing data visualization as a tool for exploratory research and argue for a need to create a greater role for it in the process of scientific discovery. For example, Carneiro and Mylonakis explain how Google Flu relies on data visualization tools to predict outbreaks of influenza by tracking online search behavior and comparing it to geographical data. Similarly, Mitchell notes how Google Maps analyzes traffic patterns through data provided via real-time cell phone location to provide recommendations for travel directions. In the realm of social media, Bollen and colleagues have also demonstrated how analysis of Twitter feeds can be used to predict public sentiments.

According to Jou, the value of behavioral analytics has perhaps been most notably observed in the area of commercial marketing. The consumer marketing space has borne witness to the progress made through extracting actionable and profitable

insights from user behavioral data. For example, between recommendation search engines for Amazon and teams of data scientists for LinkedIn, behavioral analytics has allowed these companies to transform their plethora of user data into increased profits. Similarly, advertising efforts have turned toward the use of behavioral analytics to glean further insights into consumer behavior. Yamaguchi discusses several tools on which digital marketers rely that go beyond examining data from site traffic.

Nagaitis notes observations that are consistent with Jou's view of behavioral analytics' impact on marketing. According to Nagaitis, in the absence of face-to-face communication, behavioral analytics allows commercial marketers to examine e-consumers through additional lenses apart from the traditional demographic and traffic tracking. In approaching the selling process from a relationship standpoint, behavioral analytics uses data collected via web-based behavior to increase understanding of consumer motivations and goals, and fulfill their needs. Examples of these sources of data include keyword searchers, navigation paths, and click-through patterns. By inputting data from these sources into machine learning algorithms, computational social scientists are able to map human factors of consumer behavior as it unfolds during purchases. In addition, behavioral analytics can use web-based behaviors of consumers as proxies for cues typically conveyed through in-person face-to-face communication. Previous research suggests that web-based dialogs can capture rich data pointing toward behavioral cues, the analysis of which can yield highly accurate predictions comparable to data collected during face-to-face interactions. The significance of this ability to capture communication cues is reflected in marketers increased ability to speak to their consumers with greater personalization that enhances the consumer experience.

Behavioral analytics has also enjoyed increasingly widespread application in game development. El-Nasr and colleagues discuss the growing significance of assessing and uncovering insights related to player behavior, both of which have emerged as essential goals for the game industry and catapulted behavioral analytics into

a central role with commercial and academic implications for game development. A combination of evolving mobile device technology and shifting business models that focus on game distribution via online platforms has created a situation for behavioral analytics to make important contributions toward building profitable businesses.

Increasingly available data on user behavior has given rise to the use of behavioral analytic approaches to guide game development. Fields and Cotton note the premium placed in this industry on data mining techniques that decrease behavioral datasets in complexity while extracting knowledge that can drive game development. However, determining cutting-edge methods in behavioral analytics within the game industry is a challenge due to reluctance on the part of various organizations to share analytic methods. Drachen and colleagues observe a difficulty in assessing both data and analytical methods applied to data analysis in this area due to a perception that these approaches represent a form of intellectual property. Sifa further notes that to the extent that data mining, behavioral analytics, and the insights derived from these approaches provide a competitive advantage over rival organizations in an industry that already exhibits fierce competition in the entertainment landscape, organizations will not be motivated to share knowledge about these methods.

Another area receiving attention for its application of behavioral analytics is business management. Noting that while much interest in applying behavioral analytics has focused on modeling and predicting consumer experiences, Géczy and colleagues observe a potential for applying these techniques to improve employee usability of internal systems. More specifically, Géczy and colleagues describe the use of behavioral analytics as a critical first step to user-oriented management of organizational information systems through identification of relevant user characteristics. Through behavioral analytics, organizations can observe characteristics of usability and interaction with information systems and identify patterns of resource underutilization. These patterns are important in

providing implications for designing streamlined and efficient user-oriented processes and services. Behavioral analytics can also offer prospects for increasing personalization during the user experience by drawing from user information provided in user profiles. These profiles contain information about how the user interacts with the system, and the system can accordingly adjust based on clustering of users.

Despite advances made in behavioral analytics within the commercial marketing and game industries, several areas are ripe with opportunities for integrating behavioral analytics to improve performance and decision-making practices. One area that has not yet reached its full potential for capitalizing on the use of behavioral analytics is security. Although Brown reports on exploration in the use of behavioral analytics to track cross-border smuggling activity in the United Kingdom through vehicle movement, the application of these techniques under the broader umbrella of security remains understudied. Along these lines and in the context of an enormous amount of available data, Jou discusses the possibilities for implementing behavioral analytics techniques to identify insider threats posed by individuals within an organization. Inputting data from a variety of sources into behavioral analytics platforms can offer organizations an opportunity to continuously monitor users and machines for early indicators and detection of anomalies. These sources may include email data, network activity via browser activity and related behaviors, intellectual property repository behaviors related to how content is accessed or saved, end-point data showing how files are shared or accessed, and other less conventional sources such as social media or credit reports. Connecting data from various sources and aggregating them under a comprehensive data plane can provide enhanced behavioral threat detection. Through this, robust behavioral analytics can be used to extract insights into patterns of behavior consistent with an imminent threat. At the same time, the use of behavioral analytics can also measure, accumulate, verify, and correctly identify real insider threats while preventing inaccurate

classification of nonthreats. Jou concludes that the result of implementing behavioral analytics in an ethical manner can provide practical and operative intelligence while raising the question as to why implementation in this field has not occurred more quickly.

In conclusion, behavioral analytics has been previously defined as a process in which large datasets consisting of behavioral data are analyzed for the purpose of deriving insights that can serve as actionable knowledge. This definition includes three goals underlying the use of behavioral analytics, namely, to enhance organizational performance, improve decision-making, and generate insights into user behavior. Given the burgeoning presence of big data and spread of data mining techniques to analyze this data, several fields have begun to integrate behavioral analytics into their approaches for problem-solving and performance-enhancing actions. While concerns related to accuracy and ethical use of these insights remain to be addressed, behavioral analytics can present organizations and business with unprecedented opportunities to enhance business, management, and operations.

Cross-References

- [Big Data](#)
- [Business Intelligence Analytics](#)
- [Data Mining](#)
- [Data Science](#)
- [Data Scientist](#)

Further Reading

- Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proceedings of the Fifth International Association for Advancement of Artificial Intelligence Conference on Weblogs and Social Media*.
- Brown, G. M. (2007). Use of kohonen self-organizing maps and behavioral analytics to identify cross-border smuggling activity. *Proceedings of the World Congress on Engineering and Computer Science*.
- Carneiro, H. A., & Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10).
- Davenport, T., & Harris, J. (2007). *Competing on analytics: The new science of winning*. Boston: Harvard Business School Press.
- Drachen, A., Sifa, R., Bauckhage, C., & Thureau, C. (2012). Guns, swords and data: Clustering of player behavior in computer games in the wild. *Proceedings of the IEEE Computational Intelligence and Games*.
- El-Nasr, M. S., Drachen, A., & Canossa, A. (2013). *Game analytics: Maximizing the value of player data*. New York: Springer Publishers.
- Fields, T. (2011). *Social game design: Monetization methods and mechanics*. Boca Raton: Taylor & Francis.
- Fox, P., & Hendler, J. (2011). Changing the equation on scientific data visualization. *Science*, 331(6018).
- Géczy, P., Izumi, N., Shotaro, A., & Hasida, K. (2008). Toward user-centric management of organizational information systems. *Proceedings of the Knowledge Management International Conference*, Langkawi, Malaysia (pp. 282-286).
- Kohavi, R., Rothleder, N., & Simoudis, E. (2002). Emerging trends in business analytics. *Communications of the ACM*, 45(8).
- Mitchell, T. M. (2009). Computer science: Mining our reality. *Science*, 326(5960).
- Montibeller, G., & Durbach, I. (2013). Behavioral analytics: A framework for exploring judgments and choices in large data sets. Working Paper LSE OR13.137. ISSN 2041-4668.
- Negash, S., & Gray, P. (2008). *Business intelligence*. Berlin/Heidelberg: Springer.
- Sifa, R., Drachen, A., Bauckhage, C., Thureau, C., & Canossa, A. (2013). Behavior evolution in tomb raider underworld. *Proceedings of the IEEE Computational Intelligence and Games*.

Bibliometrics/Scientometrics

Staća Milojević¹ and Loet Leydesdorff²

¹Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

²Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands

“Scientometrics” and “bibliometrics” can be used interchangeably as the name of a scientific field at

the interface between library and information science (LIS), on the one side, and the sociology of science, on the other. On the applied side, this field is well known for the analysis and development of evaluative indicators such as the journal impact factor, *h*-index, and university ranking.

The term “bibliometrics” was coined by Pritchard (1969), to describe research that utilizes mathematical and statistical methods to study written records. “Scientometrics” emerged as the quantitative study of science in the 1970s (Elkana et al. 1978), alongside with the development of citation databases (indexes) by Eugene Garfield (1955). The pioneering work in this area by the historian of science Derek de Solla Price (e.g., 1963, 1965) proposed studying the sciences as networks of documents. The citation indexes provided the measurement tools for testing hypotheses in the Mertonian sociology of science, in which one focuses on the questions of stratification and the development of scientific fields using, for example, co-citation analysis (e.g., Mullins 1973).

For example, one has also focused on questions related to the social structures that lead to advancement of science. While some researchers study larger units, such as scientific fields or disciplines, others were interested in identifying the roles played by elite and non-elite scientists. This latter question is known as the Ortega Hypothesis after the Spanish philosopher Ortega y Gasset (1932) who proposed that non-elite scientists also play a major role in the advancement of science. However, Newton’s aphorism that leading scientists “stand on the shoulders of giants” provides an alternative view of an elite structure operating as a relatively independent layer (Bornmann et al. 2010; cf. Merton 1965). In her book *The New Invisible College*, Caroline Wagner (2008) argued that international co-authorship relations have added a new layer in knowledge production during the past decades, but not in all disciplines to the same extent. In general, understanding the processes of knowledge creation is of paramount importance not only for understanding science but also for making informed decisions about the allocation of resources.

The use of citation analysis in research evaluation followed on the applied side of the field. The US National Science Board launched the biannual Science Indicator series in 1972. This line of research has grown significantly prompting a whole field of “evaluative bibliometrics” (Narin 1976). The interest in developing useful indicators has been advanced by the Organisation for Economic Co-operation and Development (OECD) and their *Frascati Manual* (OECD [1962] 2015) for the measurement of scientific and technical activities and the *Oslo Manual* (OECD [1972] 2018) for the measurement of innovations. Patents thus emerged as a useful data source to study the process of knowledge diffusion and the transfer of knowledge between science and technology (Price 1984; Rosenberg 1982). The analysis of patents has helped to make major advances in understanding processes of innovation (Jaffe and Trajtenberg 2002), and patent statistics has become a field in itself, but with strong connections to scientometrics.

While researchers have been developing ever more sophisticated indicators, some of the earlier measures became widely used in research management and policy making. The best-known of these are journal impact factors, university rankings, and the *h*-index. The impact factor, defined as a 2-year moving citation average at the level of journals, was proposed by Eugene Garfield as a means to assess journal quality for potential inclusion in the database (Garfield and Sher 1963; Garfield 1972). However, it is not warranted to use this measure for the assessment of individual publications or individual scholars for the purposes of funding and promotion given the skewness of citation distributions. Instead, citation data can also be analyzed using nonparametric statistics (e.g., percentiles) after proper normalization for a field. However, the normalization of publication and citation counts for different fields of science has remained a hitherto insufficiently solved problem.

The *h*-index (Hirsch 2005) is probably the most widely used metrics for assessing the impact of individual authors and has been praised as a simple-to-understand indicator

capable of capturing both productivity and impact. However, Waltman and Van Eck (2012) have shown that the *h*-index is mathematically inconsistent. The publication of the first Academic Ranking of World Universities (ARWU) of the Shanghai Jiao Tong University in 2004 (Shin et al. 2011) has further enhanced the attention to evaluation and ranking.

In the 2000s, new citation indexes were created, including Scopus, Google Scholar, and Microsoft Academic. As scholarly communication diversified and expanded to the web, new sources for gathering alternative metric (“altmetric”) data became available as well (e.g., Mendeley) leading to the development of indicators capable of capturing the changes in research practices, such as bookmarking, citing, reading, and sharing.

More recently, massive investments in knowledge infrastructures and the increased awareness of the importance of data sharing (Borgman 2015) have led to the attempts to provide proper incentives and recognition for authors who make their data available to a wider community. Data citation, for example, prompted the creation of the *Data Citation Index*, and a body of research focused on better understanding of data life cycles within the sciences.

In summary, the analysis of written records in quantitative science studies has significantly advanced the knowledge about the structures and dynamics of science and the process of innovation (Leydesdorff 1995). The field of scientometrics has developed into a scientific community with an intellectual core of research agendas (Milojević and Leydesdorff 2013; Wouters and Leydesdorff 1994). In addition to intensified efforts to improve indicators, the increased availability of data sources has brought a renewed interest in fundamental questions, such as theory of citation, classifications of the sciences, and the nature of collaboration across disciplines and in university-industry-government relations, and brought about great advances in the mapping of science, technology, and innovation.

The rapid increase in the number, size, and quality of data sources that are widely available and amenable to automatic processing,

together with major advances in analysis techniques, such as network analysis, machine learning, and natural language processing, holds a great potential for using these tools and techniques not only to advance scientific knowledge but also as a basis for improving decision-making when it comes to the allocation of resources.

Cross-References

- [Scientometrics](#)
- [Bibliometrics/Scientometrics](#)

References

- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge: The MIT Press.
- Bornmann, L., De Moya-Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS One*, 5(10), e13327.
- de Solla Price, D. J. (1963). *Little science, big science*. New York: Columbia University Press.
- de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(30), 510–515.
- de Solla Price, D. J. (1984). The science/technology relationship, the craft of experimental science, and policy for the improvement of high technology innovation. *Research Policy*, 13(1), 3–20.
- Elkana, Y., Lederberg, J., Merton, R. K., Thackray, A., & Zuckerman, H. (Eds.). (1978). *Toward a metric of science: The advent of science indicators*. New York: Wiley.
- Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159), 108–111.
- Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178, 471–479.
- Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14(3), 195–201.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS*, 102(46), 16569–16572.
- Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. Cambridge: The MIT Press.
- Leydesdorff, L. (1995). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications*. Leiden: DSWO Press, Leiden University.

- Merton, R. K. (1965). *On the shoulders of giants: A Shandean postscript*. New York: The Free Press.
- Milojević, S., & Leydesdorff, L. (2013). Information metrics (iMetrics): A research specialty with a socio-cognitive identity? *Scientometrics*, 95(1), 141–157.
- Mullins, N. C. (1973). *Theories and theory groups in contemporary American sociology*. New York: Harper & Row.
- Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill: Computer Horizons.
- OECD. (2015 [1962]). *The measurement of scientific and technical activities: "Frascati Manual"*. Paris: OECD. Available at https://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2015_9789264239012-en.
- OECD/EuroStat (2018 [1972]). Proposed guidelines for collecting and interpreting innovation data, "Oslo manual". Paris: OECD. Available at https://www.oecd-ilibrary.org/science-and-technology/oslo-manual-2018_9789264304604-en.
- Ortega y Gasset, J. (1932). *The revolt of the masses*. New York: Norton.
- Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25, 348–349.
- Rosenberg, N. (1982). *Inside the black box: Technology and economics*. Cambridge: Cambridge University Press.
- Shin, J. C., Toutkoushian, R. K., & Teichler, U. (2011). *University rankings: Theoretical basis, methodology and impacts on global higher education*. Dordrecht: Springer.
- Wagner, C. S. (2008). *The new invisible college: Science for development*. Washington, DC: Brookings Institution Press.
- Waltman, L., & Van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2), 406–415.
- Wouters, P., & Leydesdorff, L. (1994). Has Price's dream come true: Is scientometrics a hard science? *Scientometrics*, 31(2), 193–222.

Big Data

- Business Intelligence Analytics
- Data Integration
- Data Provenance
- NoSQL (Not Structured Query Language)

Big Data Analytics

- NoSQL (Not Structured Query Language)

Big Data and Theory

Wolfgang Maass¹, Jeffrey Parsons², Sandeep Purao³, Alirio Rosales⁴, Veda C. Storey⁵ and Carson C. Woo⁴

¹Saarland University, Saarbrücken, Germany

²Memorial University of Newfoundland, St. John's, Canada

³Bentley University, Waltham, USA

⁴University of British Columbia, Vancouver, Canada

⁵J Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA

The necessity of grappling with Big Data, and the desirability of unlocking the information hidden within it, is now a key theme in all the sciences – arguably the key scientific theme of our times. (Diebold 2012)

Introduction

Big data is the buzzword du jour in diverse fields in the natural, life, social, and applied sciences, including physics (Legger 2014), biology (Howe et al. 2008), medicine (Collins and Varmus 2015), economics (Diebold 2012), and management (McAfee and Brynjolfsson 2012; Gupta and George 2016). The traditional Vs of big data – volume, variety, and velocity – reflect the unparalleled quantity, diversity, and immediacy of data generated by sensing, measuring, and social computing technologies. The result has been significant new research opportunities, as well as unique challenges. Computer and information scientists have responded by developing tools and techniques for big data analytics, intended to discover patterns (statistical regularities among variables) in massive data sets (Fukunaga 2013), reconcile the variety in diverse sources of data (Halevy et al. 2009), and manage data generated at a high velocity.

With the success of these tools and techniques, some have proclaimed the “end of theory,” arguing that “the data deluge makes the scientific

method obsolete” (Anderson 2008) and that any question can now be answered by data analysis (Halevy et al. 2009; Baesens et al. 2016). This position has led to a radical rethinking of scientific practice, as well as an assessment of the impact of big data research in specific disciplines, such as astronomy (Pankratius and Mattmann 2014) and biology (Shaffer and Purugganan 2013).

However, a primary focus on statistical pattern finding in big data has limited potential to advance science because the extracted patterns can be misleading or reveal only idiosyncratic relationships (Bentley et al. 2014). Research based on big data analytics should be complemented by theoretical principles that evaluate which patterns are meaningful. Big data and big theory should complement each other. Researchers, thus, need to integrate theory and big data analytics for conducting science in the era of big data (Rai 2016).

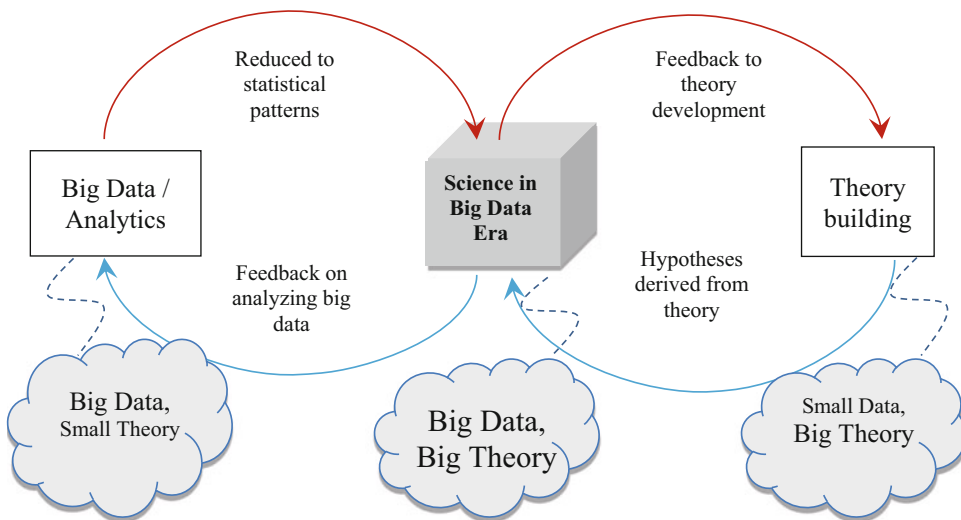
Framework for Science in the Era of Big Data

Big data analytics and domain theories should have complementary roles as scientific practice moves toward a desirable future that combines “big data” with “big theory.” This “big data, big

theory” is in contrast with the traditional scientific focus on “small data, big theory,” going beyond data-driven emphasis on “big data, small theory,” by explicating interactions between big data analytics and theory building.

Figure 1 presents a framework for *science in the era of big data* that represents these interactions. The intent of the framework is to identify possibilities that researchers can use to position their work, thereby encouraging closer interactions between research communities engaged in big data analytics versus theory-driven research. The value of the framework can be explored by examining scientific practice, which has primarily been driven by the cost and effort required for data collection and analysis.

Work in natural sciences has focused on developing or testing theory with relatively small data sets, often due to the cost of experimental design and data collection. As recently as the first decade of the twenty-first century, population genetic models were based only upon the analysis of one or two genes, but now evolutionary biologists can use data sets at the scale of the full genome of an increasing number of species (Wray 2010). Analogous examples exist in other fields, including sociology and management (Schilpzand et al. 2014). This mode of research emphasizes a tight link between big data analytics and theory



Big Data and Theory, Fig. 1 Framework for science in the era of big data

building and testing, exemplifying a mode of research we call “small data, big theory.”

Sensor technologies and massive computing power have transformed data collection and analysis by reducing effort and cost. Scientists can now extract statistical patterns from very large data sets with advanced analytical techniques (e.g., Dhar 2013). Biomedical scientists can analyze full genomes (International Human Genome Sequencing Consortium 2004; ENCODE Project Consortium 2012). Likewise, astronomy is becoming a computationally intensive field due to an “exciting evolution from an era of scarce scientific data to an era of overabundant data” (Shaffer and Purugganan 2013). Research in these domains is being transformed with the use of big data techniques that may have little or no connection to prior theories in the scientific discipline. This practice exemplifies a mode of research we call “big data, small theory.”

This emphasis on big data analytics risks severing the connection between data and theory, threatening our ability to understand and interpret extracted statistical patterns. Overcoming this threat requires purposeful interactions between theory development and data collection and analysis. The framework highlights these interactions via labeled arrows.

We are already beginning to witness such interactions. Population geneticists, for example, can delve deeper into our evolutionary past by postulating the genetic structure of extinct and ancestral populations and investigating them with the help of novel sequencing technologies and other methods of data analysis (Wall and Slatkin 2012). A new field of “paleopopulation genetics” was not possible without proper integration of big data and theory. In astronomy, statistical patterns can easily be extracted from large data sets, although theory is required to interpret them properly (Shaffer and Purugganan 2013). The standard model, a fundamental theory in particle physics, places requirements on energies needed for producing experimental conditions for the Higgs boson (Aad et al. 2012). Based upon these theory-derived requirements, scientists have verified the theoretical prediction of the Higgs boson by

analyzing big data created by the Large Hadron Collider.

Core research areas in computer science are affected by big data analytics. For instance, computer vision witnessed a major shift as the concept of deep learning (Krizhevsky et al. 2012) significantly improved the success rates of many applications such as facial recognition. It also changed research to an algorithmic understanding of computer vision. Image recognition now obtains results that are getting close to that of humans (Sermanet et al. 2013), which was not feasible with prior declarative theories. When images can be classified with humanlike accuracy, even better scientific questions can be posed, such as “what really is vision?” by generating procedural theories that replicate and explain how the human brain operates (LeCun et al. 2015). In this sense, research is moving from “what is built” to “how to build.”

These examples highlight a desirable mode of research, “big data, big theory.” This form of research includes extracting statistical patterns from large, and often heterogeneous, data sets. Pattern extraction is not a stand-alone activity, but rather one that shapes, and is shaped by, theory building and testing.

Application of Framework

The framework for science in the era of big data in Fig. 1 depicts and promotes interdisciplinary interactions between researchers in the big data analytics field (L.H.S.) with those in disciplines or domains related to various sciences (R.H.S.).

The top left arrow indicates that a data scientist has the capability to reduce big data to statistical patterns using analytical techniques, as in the identification of homologous sequences from evolutionary genomics (Saitou 2013).

The top right arrow shows that statistical patterns can suggest novel theoretical explanations to domain scientists. These patterns may extend theory to new domains or reveal inconsistencies in current theory. Without integration with theory, statistical patterns may remain as merely curious facts (Bentley et al. 2014). In some situations, big

data may significantly increase empirical support for existing theoretical predictions. In other cases, big data may simplify theory testing (e.g., by modifying measurement parameters), facilitate theory refinement (e.g., based upon more or new data), or radically extend the scope of a theory. An example is paleopopulation genetics (Wall and Slatkin 2012), which has made possible studies of extinct populations.

The bottom right arrow indicates that a domain scientist can identify specific data that may need to be acquired, by revealing gaps in testing existing theories. This leads to the bottom left arrow where data scientists can close a gap by extracting, cleaning, integrating, and analyzing relevant data. Doing so might also reveal alternative perspectives on how to manipulate, analyze, or synthesize existing (big) data. For example, Metcalfe's law could only be properly tested when large amounts of data became available (e.g., membership growth numbers from Facebook) (Metcalfe 2013).

Another example of "big data, big theory" is the emerging discipline of astroinformatics. It would be incorrect to view computing in astronomy as applied computer science. Clearly, computer science impacts astronomy, but computer scientists do not have effective techniques that can be easily adapted to astronomy (Shaffer and Purugganan 2013). Through interaction with astronomers, techniques are created and evolve. This interdisciplinary integration highlights a crucial aspect of the changing nature of scientific practice. Considering big data and theory-driven research as complementary endeavors can produce outcomes not achievable by considering either in isolation.

Conclusion

Computing technologies provide an exciting opportunity for a new mode of research (big data, big theory), as the scientific community moves from a time of "data poverty" to "data wealth." The science in the era of big data framework provides both data and domain scientists with an understanding of how to

position themselves at the desired intersection of big data and big theory, to realize the potential for unprecedented scientific progress.

Reasonable actions that researchers should consider from the data analytics perspective include: (1) using larger and more complete data sets (e.g., physics, biology, and medicine); (2) increasing computational capabilities (e.g., astronomy); (3) mining heterogeneous data sets for predictive analytics, text mining, and sentiment (e.g., business applications); (4) adopting new machine learning techniques (e.g., computer vision); and (5) generating new, and novel, questions. From a theoretical perspective, researchers should consider: (1) what impactful theoretical questions can now be addressed that could not be answered using the traditional "big theory, small data" approach; (2) how interpretability of patterns extracted can be supported by or drive theory development; and (3) how theoretical concepts can be mapped onto available data variables and vice versa.

Minimally, the framework can enable scientists to reflect on their practices and better understand why theory remains essential in the era of big data. An extreme interpretation of the framework is a reconceptualization of the scientific endeavor itself; indeed, one that recognizes the synergy between big data and theory building as intrinsic to future science. The framework has been illustrated in several domains to demonstrate its applicability across disciplines.

Further Reading

- Aad, G., et al. (2012). Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1), 1–29.
- Anderson, C. (2008). The end of theory. *Wired Magazine*, 16(7), 16–07.
- Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2016). Transformational issues of big data and analytics in networked business. *MIS Quarterly*, 40(4), 807–818.
- Bentley, R. A., O'Brien, M. J., & Brock, W. A. (2014). Mapping collective behavior in the big-data era. *Behavioral and Brain Sciences*, 37(01), 63–76.

- Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9), 793–795.
- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64–73.
- Diebold, F. X. (2012). *On the origin (s) and development of the term 'Big Data'* (PIER working paper). Philadelphia: PIER.
- ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.
- Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Academic press, Cambridge, MA.
- Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information Management*, 53(8), 1049–1064.
- Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12.
- Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... & Twigger, S. (2008). Big data: The future of biocuration. *Nature*, 455(7209), 47–50.
- International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931–945.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105), Curran Associates, Red Hook, NY.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Legger, F. (2014). The ATLAS distributed analysis system. *Journal of Physics: Conference Series*, 513(3):032053.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10), 60–68.
- Metcalfe, B. (2013). Metcalfe's law after 40 years of ethernet. *Computer*, 46(12), 26–31.
- Pankratius, V., & Mattmann, C. (2014). Computing in astronomy: To see the unseen. *Computer*, 9(47), 23–25.
- Rai, A. (2016). Synergies between big data and theory. *MIS Quarterly*, 40(2), iii–ix.
- Saitou, N. (2013). *Introduction to evolutionary genomics*. London: Springer.
- Schilpzand, P., Hekman, D. R., & Mitchell, T. R. (2014). An inductively generated typology and process model of workplace courage. *Organization Science*, 26(1), 52–77.
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.
- Shaffer, H. B., & Purugganan, M. D. (2013). Introduction to theme “Genomics in Ecology, Evolution, and Systematics”. *Annual Review of Ecology, Evolution, and Systematics*, 44, 1–4.
- Wall, J. D., & Slatkin, M. (2012). Paleopopulation genetics. *Annual Review of Genetics*, 46, 635–649.
- Wray, G. A. (2010). Integrating genomics into evolutionary theory. In M. Pigliucci & G. B. Muller (Eds.), 2010 *Evolution: The extended synthesis* (pp. 97–116). Cambridge, MA: MIT Press.

Big Data Concept

Connie L. McNeely and Laurie A. Schintler
George Mason University, Fairfax, VA, USA

Big data is one of the most critical features of today's increasingly digitized and expanding information society. While the term “big data” has been invoked in various ways in relation to different stakeholders, groups, and applications, its definition has been a matter of some debate, changing over time and focus. However, despite a lack of consistent definition, typical references point to the collection, management, and analysis of massive amounts of data, with some general agreement signaling the size of datasets as the principal defining factor. As such, big data can be conceptualized as an analytical space marked by encompassing processes and technologies that can be employed across a wide range of domains and applications. Big data are derived from various sources – including sensors, observatories, satellites, the World Wide Web, mobile devices, crowdsourcing mechanisms, and so on – to the extent that attention is required to both instrumental and intrinsic aspects of big data to understand their meanings and roles in different circumstances, sectors, and contexts. This means considering conceptual delineations and analytical uses relative to issues of validity, credibility, applicability, and broader implications for society today and in the future.

Conceptual Dimensions

The explosion of big data references the breadth and depth of the phenomenon in and of itself as a core operational feature of society relative to how we understand and use it in that regard. Big data is

a multidimensional concept which, despite different approaches, largely had been interpreted initially according to the “3 Vs”: volume, variety, and velocity. *Volume* refers to the increasing amount of data; *variety* refers to the complexity and range of data types and sources; and *velocity* refers to the speed of data, particularly the rate of data creation and availability. That is, big data generally refers to massive volumes of data characterized by variety that reflects the heterogeneity and types of structured and unstructured data collected and the velocity at which the data are acquired and made available for analysis. These dimensions together constitute a basic conceptual model for describing big data.

However, beyond the initial basic model, two additional Vs – variability and veracity – have been noted, to the extent that reference to the “5 Vs” became common. *Variability* is reflected in inconsistencies in data flows that attend the variety and complexity that mark big data. *Veracity* refers to the quality and reliability of the data, i.e., indicating data integrity and the extent to which the data can be trusted for analytical and decision-making purposes. Veracity is of special note given that big data often are compromised by incompleteness, redundancy, bias, noise, and other imperfections. Accordingly, methods of data verification and validation are especially relevant in this regard. Following these lines, a sixth V – vulnerability – has been recognized as a fundamental and encompassing characteristic of big data. *Vulnerability* is an integrated notion that speaks to security and privacy challenges posed by the vast amounts, range of sources and formats, and the transfer and distribution of big data, with broad social implications. Also, a seventh V – value – is typically discussed in this regard as yet another consideration. *Value*, as a basic descriptive dimension, highlights the capacity for adding and extracting value from big data. Thus, allusions to the “7 Vs” – volume, variety, velocity, variability, veracity, vulnerability, and value – have been increasingly invoked as broadly indicative of big data and are now considered the principal determinant features that mark related conceptual delineations. (Note that, on occasion, three other Vs also have been included – volatility,

validity, and viscosity – although these generally are encompassed in the other Vs. They typically are raised as specific points rather than in overall reference to big data.)

Big data contains disparate formats, structures, semantics, granularity, and so on, along with other dimensions related to exhaustivity, identification, relationality, extensionality, and scalability. Specifically, big data can be described in terms of how well it captures a complete system or population; its resolution and ability to be indexed; the ease with which it can be reduced, expanded, or integrated; and its potential to expand in size rapidly. Big data also can be delineated by spatial and/or temporal features and resolution. The idea that things can be learned from a large body of data that cannot be comprehended from smaller amounts also links big data to notions of complexity, such that complex structures, behaviors, and permutations of datasets are basic considerations in labeling data as big. Big data need not incorporate all of the same characteristics and, in fact, few big data sets possess all possible dimensions, to the effect that there can be multiple forms of big data.

Conceptual Sources and Allocations

Massive datasets derive from a variety of sources. For example, the digitization of large collections of documents has given rise to massive corpora and archives of unstructured data. Additionally, social media, crowdsourcing platforms, e-commerce, and other web-based sources are contributing to a vast and complex collection of information on social and economic exchanges and interactions among people, places, and organizations from moment-to-moment around the world. Satellites, drones, telescopes, and other modes of surveillance are collecting massive amounts of information on the physical and human-made environment (buildings, nightlights, land use cover, meteorological conditions, water quality, etc.) as well as the cosmos. Big data also has been characterized as “organic,” i.e., continuously produced and observational transaction data from the everyday behaviors of people. Along

those lines, for example, mobile devices (e.g., “smart phones”) and location acquisition technologies are producing reams of detailed information on people, animals, the world, and various phenomena. The Internet of Things (IoT), which comprises a large and growing assemblage of interconnected devices, actively monitors and processes everything from the contents of refrigerators to the operational characteristics of large-scale infrastructures. Large-scale simulations based on such data provide additional layers of data.

Big data has abundant applications and is being used across virtually all sectors of society. Some industries in particular have benefitted from big data availability and use relative to others: healthcare, banking and finance, media, retail, and energy and utilities. Beyond those, industries that are rapidly being encompassed and marked by big data include medicine, construction, and transportation. Worldwide, the industries that are investing the most in big data are banking, manufacturing, professional services, and government, with effects manifesting across all levels of analysis.

Analytical and Computational Capabilities

The generation, collection, manipulation, analysis, and use of big data can make for a number of challenges, including, for example, dealing with highly distributed data sources, tracking and validating data, confronting sampling biases and heterogeneity, working with variably formatted and structured data, ensuring data integrity and security, developing appropriately scalable and incremental algorithms, and enabling data discovery, integration, and sharing. Accordingly, the tools and techniques that are used to process – to search, aggregate, and cross-reference – massive datasets play key roles in producing, manipulating, and recognizing big data as such. Related capabilities rest on robust systems at each point along the data pipeline, and algorithms for handling related tasks, including allocating, pooling, and

coordinating data resources, are central to managing the volume, velocity, and variety of big data.

The computational strategies and technologies that are used to handle large datasets also offer a conceptual frame for understanding big data, and artificial intelligence (AI) and machine learning (ML) are being employed to make sense of the complex and massive amounts of data. Moreover, the expanding amounts, availability, and variety of data further empower and support AI and ML applications. Along with tools for processing language, images, video, and audio, AI and ML are advancing capacities to glean insights and intelligence from big data, and other technologies, such as cloud computing, are enhancing the ability to store and process big data. However, while tools and methods are available for handling the complexity of big data in general, more effective approaches are needed for greater specification in dealing with various types of big data (e.g., spatial data) and for assessing and comparing data quality, computing efficiency, and the performance of algorithms under different conditions and across different contexts.

Conclusion

Big data is one of the most pertinent and defining features of the world today and will be even more so in the future. Broadly speaking, big data refers to the collection, analysis, and use of massive amounts of digital information for decision making and operational applications.

With data expected to grow even bigger in terms of pervasiveness, scale, and value in the near future (a process that arguably has accelerated due to the pandemic-intensified growth of “life online”), big data tools and technologies are being developed to allow for the real-time processing and management of large volumes of a variety of data (e.g., generated from IoT devices) to reveal and specify trends and patterns and to indicate relationships and connections to inform relevant decision making, planning, and research.

Big data, as a term, indicates large volumes of information from a variety of sources coming at

very high speeds. However, the wide range of sources and quality of big data can lead to problems and challenges to its use. The opportunities and challenges brought on by the explosion of information require considerations of problems that occur with and because of big data. The massive size and high dimensionality of big datasets present computational challenges and problems of validation linked to not only selection bias and measurement errors, but also to spurious correlations, storage and scalability blockages, noise accumulation, and incidental endogeneity. Moreover, the bigger the data, the bigger the potential not just for its use, but, importantly, for its misuse, including ethical violations, discrimination, and bias. Accordingly, policies and basic approaches are needed to ensure that the possible benefits of big data are maximized, while the downsides are minimized. The bottom line is that the future will be affected by how big data are collected, managed, used, and understood. Data is the foundation of the digital world, and big data are and will be fundamental to determining and realizing value in that context.

Further Reading

- boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, and Society*, 15(5), 662–679.
- Economist*. (2010, 27 February). Data data everywhere: A special report on managing information. <https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf>.
- Ellingwood, J. (2016, 28 September). An introduction to big data concepts and terminology. *Digital Ocean*. <https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology>.
- Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2), 293–314.
- Frehill, L. M. (2015). Everything old is new again: The big data workforce. *Journal of the Washington Academy of Sciences*, 101(3), 49–62.
- Galov, N. (2020, 24 November). 77+ big data stats for the big future ahead | Updated 2020. <https://hostingtribunal.com/blog/big-data-stats>.
- Groves, R. (2011, 31 May). ‘Designed data’ and ‘organic data.’ *U.S. Census Bureau Director’s Blog*. <https://www.census.gov/newsroom/blogs/director/2011/05/1309.5821v1.pdf>.
- Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data and Society*, 3(1), 1–10.
- Lohr, S. (2013, 19 June). Sizing up big data, broadening beyond the internet. *New York Times*. http://bits.blogs.nytimes.com/2013/06/19/sizing-up-big-data-broadening-beyond-the-internet/?_r=0.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.
- McNeely, C. L. (2015). Workforce issues and big data analytics. *Journal of the Washington Academy of Sciences*, 101(3), 1–11.
- McNeely, C. L., & Hahm, J. (2014). The big (data) bang: Policy, prospects, and challenges. *Review of Policy Research*, 31(4), 304–310.
- Schintler, L. A. (2020). Regional policy analysis in the era of spatial big data. In Z. Chen, W. Bowen, & D. Whittington (Eds.), *Development studies in regional science* (pp. 93–109). Singapore: Springer.
- Schintler, L. A., & Chen, Z. (Eds.). (2017). *Big data for regional science*. New York: Routledge.
- Smartym Pro. (2020). How to protect big data? The key big data security challenges. <https://smartym.pro/blog/how-to-protect-big-data-the-main-big-data-security-challenges>.
- Ward, J. S., & Barker, A. (2013). Undefined by data: A survey of big data definitions. <http://arxiv.org/pdf/1309.5821v1.pdf>.

Big Data Hubs and Spokes

- [Big Data Research and Development Initiative \(Federal, U.S.\)](#)

Big Data Integration Tools

- [Data Integration](#)

Big Data Literacy

Padmanabhan Seshaiyer and Connie L. McNeely
George Mason University, Fairfax, VA, USA

Over the last decade, the need to understand big data has emerged as an increasingly important

area that is making an impact on the least to the most advanced enterprises and societies in the world. Whether it is about analyzing terabytes to petabytes of data or recognizing and revealing or predicting patterns of behavior and interactions, the growth of big data has emphasized the pressing need to develop next generation data scientists who can anticipate user needs and develop “intelligent services” to address business, academic, and government challenges. This would require the engagement of big data proficient students, faculty, and professionals who will help to bridge the big data to knowledge gap in communities and organizations at local, national, and global levels. However, the sheer pervasiveness of big data also makes clear the need for the population in general to have a better understanding of the collection and uses of big data as it affects them directly and indirectly, within and across sectors and circumstances. Especially relevant for addressing community needs in regard to sustainability and development, in public, personal, and professional milieus, at least a basic understanding of big data is increasingly required for life in the growing information society. In other words, there is a general call for “big data literacy,” and in this sense, *big data literacy refers to the basic empowerment of individuals and groups to better understand, use, and apply big data assets for effective decision making and societal participation and contribution*. So, how is a big data literate populace created, with the ability to address related challenges in real-world settings?

Reality reflects a critical disconnect between big data understanding and awareness cutting across different groups and communities. Related possibilities and expectations must be assessed accordingly, especially since education is linked to financial, social, psychological, and cognitive conditions that promote or hinder literacy development. Underscoring the content dimension of literacy asymmetries, even if people have technological access (e.g., access to computers and the Internet), they may be limited by knowledge gaps that separate them from any meaningful understanding of big data roles, uses, and effects.

In education, acquiring big data acumen is based on exposure to material from multiple

areas of study – most notably, mathematical, statistical, and computational foundations. Particularly in reference to developing a data science ready workforce, big data literacy is a complex undertaking, involving knowledge of data acquisition, modeling, management and curation, visualization, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving (cf. NASEM 2018). However, given the need not only for a big data literate workforce, but also for a general population with a basic grasp of big data collection and uses, it is important to look beyond training individuals in data science analytics and skills to a more broad-based competency in big data.

In today’s increasingly digitized real-world contexts, the need to understand, use, and apply big data to address everyday challenges is expected. An example can be found in the contributions of big data to critical questions of sustainability and development (Sachs 2012). For instance, big data can help provide real-time information on income levels via spending patterns on mobile devices. Another example is tracking access to clean water through data collected from sensors connected to water pumps. Big data also is used in analysis of social media in revealing public opinions on various issues, such as effective governance or human rights. Examples such as these suggest the importance of the need for big data literacy in solving global challenges.

Contexts can be engaged to help build a meaningful map to data competency. Frameworks such as design thinking or system thinking provide user-centered approaches to big data problem solving. Identifying and defining problems, providing guidelines for accessible solutions, and assessing implementation feasibility are integral to building big data literacy and competency. Such frameworks can be used to decompose big problems into small problems (big data to small data) and design adaptive and effective strategies. Central to this process – and key for understanding – are capacities for big data interpretation, planning, and application for making decisions and navigating life in a digitized environment.

Big data literacy and related skills concern capacities for effective and efficient uses of data resources, encompassing awareness of what is possible, knowing how to find applicable information, being able to assess content validity and perform related tasks, and engaging and managing relevant knowledge. Defined in terms of understanding, use, and application in real-world contexts, big data literacy is critical for sustainability and development for individuals and communities today and in the future. Having said that, a cautionary note is also relevant for big data literacy, particularly regarding the ethical implications and impacts – including bias and representation – of big data collection and uses through artifacts such as images, social media, videos, etc. (Newton et al. 2005). For example, questions of surveillance, monitoring, and privacy are particularly relevant in terms of big data collection and uses and related effects (Xu and Jia 2015). Awareness of these kinds of issues is central to developing big data literacy.

Cross-References

- [Data Scientist](#)
- [Digital Literacy](#)
- [Ethics](#)
- [Privacy](#)

Further Reading

- National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. Washington, DC: National Academies Press.
- Newton, E. M., Sweeney, L., & Malin, B. (2005). Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2), 232–243.
- Sachs, J. D. (2012). From millennium development goals to sustainable development goals. *The Lancet*, 379 (9832), 2206–2211.
- Xu, H., & Jia, H. (2015). Privacy in a networked world: New challenges and opportunities for privacy research. *Journal of the Washington Academy of Sciences*, 101 (3), 73–84.

Big Data Quality

Subash Thota

Synectics for Management Decisions, Inc.,
Arlington, VA, USA

Introduction

Data is the most valuable asset for any organization. Yet in today's world of big and unstructured data, more information is generated than can be collected and properly analyzed. The onslaught of data presents obstacles to create data-driven decisions. *Data quality* is an essential characteristic of data that determines the reliability of data for making decisions in any organization or business. Errors in data can cost a company millions of dollars, alienate customers, and make implementing new strategies difficult or impossible (Redman 1995).

In practically every business instance, project failures and cost overruns are due to fundamental misunderstanding about the data quality that is essential to the initiative. A global data management survey by PricewaterhouseCoopers of 600 companies across the USA, Australia, and Britain showed that 75% of reported significant problems were a result of data quality issues, with 33% of those saying the problems resulted in delays in getting new business intelligence (BI) systems running or in having to scrap them altogether (Capehart and Capehart 2005). The importance and complexity related to data and its quality compounds incrementally and could potentially challenge the very growth of the business that acquired the data. This paper is intended to showcase challenges related to data quality and approaches to mitigating data quality issues.

Data Defined

Data is “... language, mathematical or other symbolic surrogates which are generally agreed upon to represent people, objects, events and concepts” (Liebenau and Backhouse 1990). Vayghan et al.

(2007) argued that most enterprises deal with three types of data: master data, transactional data, and historical data. *Master data* are the core data entities of the enterprise, i.e., customers, products, employees, vendors, suppliers, etc. *Transactional data* describe an event or transaction in an organization, such as sales orders, invoices, payments, claims, deliveries, and storage records. Transactional data is time bound and changes to historical data once the transaction has ended. *Historical data* contain facts, as of certain point in time (e.g., database snapshots), and version information.

Data Quality

Data quality is the capability of data to fulfill and satisfy the stated business, framework, system and technical requirements of an enterprise. A classic definition of data quality is “fitness for use,” or more specifically, the extent to which some data successfully serve the purposes of the user (Tayi and Ballou 1998; Cappiello et al. 2003; Lederman et al. 2003; Watts et al. 2009).

To be able to correlate data quality issues to business impacts, we must be able to both classify our data quality expectations as well as our business impact criteria. In order to do that, it is valuable to understand these common data quality dimensions (Loshin 2006):

- **Completeness:** Is all the requisite information available? Are data values missing, or in an unusable state? In some cases, missing data is irrelevant, but when the information that is missing is critical to a specific business process, completeness becomes an issue.
- **Conformity:** Are there expectations that data values conform to specified formats? If so, do all the values conform to those formats? Maintaining conformance to specific formats is important in data representation, presentation, aggregate reporting, search, and establishing key relationships.
- **Consistency:** Do distinct data instances provide conflicting information about the same underlying data object? Are values consistent across data sets? Do interdependent attributes

always appropriately reflect their expected consistency? Inconsistency between data values plagues organizations attempting to reconcile different systems and applications.

- **Accuracy:** Do data objects accurately represent the “real-world” values they are expected to model? Incorrect spellings of products, personal names or addresses, and even untimely or not current data can impact operational and analytical applications.
- **Duplication:** Are there multiple, unnecessary representations of the same data objects within your data set? The inability to maintain a single representation for each entity across your systems poses numerous vulnerabilities and risks.
- **Integrity:** What data is missing important relationship linkages? The inability to link related records together may actually introduce duplication across your systems. Not only that, as more value is derived from analyzing connectivity and relationships, the inability to link related data instance together impedes this valuable analysis.

Causes and Consequences

The “Big Data” era comes with new challenges for data quality management. Beyond volume, velocity, and variety lies the importance of the fourth “V” of big data: *veracity*. Veracity refers to the trustworthiness of the data. Due to the sheer volume and velocity of some data, one needs to embrace the reality that when data is extracted from multiple datasets at a fast and furious clip, determining the semantics of the data – and understanding correlations between attributes – becomes of critical importance.

Companies that manage their data effectively are able to achieve a competitive advantage in the marketplace (Sellar 1999). On the other hand, bad data can put a company at a competitive disadvantage comments (Greengard 1998). It is therefore important to understand some of the causes of bad data quality:

- Lack of data governance standards or validation checks.

- Data conversion usually involves transfer of data from an existing data source to a new database.
- Increasing complexity of data integration and enterprise architecture.
- Unreliable and inaccurate sources of information.
- Mergers and acquisitions between companies.
- Manual data entry errors.
- Upgrades of infrastructure systems.
- Multidivisional or line-of-business usage of data.
- Misuse of data for purposes different from the capture reason.

Different people performing the same tasks have a different understanding of the data being processed, which leads to inconsistent data making its way into the source systems. Poor data quality is a primary reason for 40% of all business initiatives failing to achieve their targeted benefits (Friedman and Smith 2011). Marsh (2005) summarizes consequences in one of his article:

- Eighty-eight percent of all data integration projects either fail completely or significantly overrun their budgets.
- Seventy-five percent of organizations have identified costs stemming from dirty data.
- Thirty-three percent of organizations have delayed or canceled new IT systems because of poor data.
- \$611B per year is lost in the USA to poorly targeted bulk mailings and staff overheads.
- According to Gartner, bad data is the number one cause of customer-relationship management (CRM) system failure.
- Less than 50% of companies claim to be very confident in the quality of their data.
- Business intelligence (BI) projects often fail due to dirty data, so it is imperative that BI-based business decisions are based on clean data.
- Only 15% of companies are very confident in the quality of external data supplied to them.
- Customer data typically degenerates at 2% per month or 25% annually.

To Marsh, organizations typically overestimate the quality of their data and underestimate the cost

of data errors. Business processes, customer expectations, source systems and compliance rules are constantly changing – and data quality management systems must reflect this. Vast amounts of time and money are spent on custom coding and “firefighting” to dampen an immediate crisis rather than dealing with the long-term problems that bad data can present to an organization.

Data Quality: Approaches

Due to the large variety of sources from which data is collected and integrated, for its sheer volume and changing nature, it is impossible to manually specify data quality rules. Below are a few approaches to mitigating data quality issues:

1. Enterprise Focus and Discipline

Enterprises should be more focused and engaged toward data quality issues; views toward data cleansing must evolve. Clearly defining roles and outlining the authority, accountability and responsibility for decisions regarding enterprise data assets provides the necessary framework for resolving conflicts and driving a business forward as the data-driven organization matures. Data quality programs are most efficient and effective when they are implemented in a structured, governed environment.

2. Implementing MDM and SOA

The goal of a master data management (MDM) solution is to provide a single source of truth of data, thus providing a reliable foundation for that data across the organization. This prevents business users across an organization from using different versions of the same data. Another approach of big data and big data governance is the deployment of cloud-based models and software-oriented architecture (SOA). SOA enables the tasks associated with a data quality program to be deployed as a set of services that can be called dynamically by applications. This allows business rules for data quality enforcement to be moved outside of applications and applied universally at

a business process level. These services can either be called proactively by applications as data is entered into an application system, or by batch after the data has been created.

3. Implementing Data Standardization and Data Enrichment

Data standardization usually covers reformatting of user-entered data without any loss of information or enrichment of information. Such solutions are most suitable for applications that integrate data. Data enrichment covers the reformatting of data with additional enrichment or addition of useful referential and analytical information.

Data Quality: Methodology in Profiling

Data profiling provides a proactive way to manage and comprehend an organization's data. Data profiling is explicitly about discovering and reviewing the underlying data available to determine the characteristics, patterns, and essential statistics about the data. Data profiling is an important diagnostic phase that furnishes quantifiable and tangible facts about the strength of the organization's data. These facts not only help in establishing what data is available in the organization but also how accurate, valid, and usable the data is. Data profiling covers numerous techniques and processes:

- **Data Consistency:** This is the fidelity or integrity of the data within data structures or interfaces.
 - **Data Adherence:** This is a measure of compliance or adherence of the data to the intended standards or logical rules that govern the storage or interpretation of data.
 - **Data Duplicity:** This is a measure of duplicates records or fields in the system that can be consolidated to reduce the maintenance costs and efficiency of the system storage processes.
 - **Data Completeness:** This is a measure of the correspondence between the real world and the specified dataset.
- In assessing a dataset for veracity, it is important to answer core questions about it:
- Do the patterns of the data match expected patterns?
 - Do the data adhere to appropriate uniqueness and null value rules?
 - Are the data complete?
 - Are they accurate?
 - Do they contain information that is easily understood and unambiguous?
 - Do the data adhere to specified required key relationships across columns and tables?
 - Are there inferred relationships across columns, tables, or databases?
 - Are there redundant data?
- Data in an enterprise is often derived from different sources, resulting in data inconsistencies and nonstandard data. Data profiling helps analysts dig deeper to look more closely at each of the individual data elements and establish which data values are inaccurate, incomplete, or ambiguous. Data profiling allows analysts to link data in disparate applications based on their relationships to each other or to a new application being developed. Different pieces of relevant data spread across many individual data stores make it difficult to develop a complete understanding of an enterprise's data. Therefore, data profiling helps one understand how data sources interact with other data sources.
- **Data Ancestry:** This covers the lineage of the dataset. It describes the source from which the data is acquired or derived and the method of acquisition.
 - **Data Accuracy:** This is the closeness of the attribute data associated with an object or feature, to the true value. It is usually recorded as the percentage correctness for each topic or attribute.
 - **Data Latency:** This is the level at which the data is current or accurate to date. This can be measured by having appropriate data reconciliation procedures to gauge any unintended delays in acquiring the data due to technical issues.

Metadata

Metadata is used to describe the characteristics of a data field in a file or a table and contains information that indicates the data type, the field length, whether the data should be unique, and if a field can be missing or null. Pattern matching determines if the data values in a field are in the likely format. Basic statistics about data such as minimum and maximum values, mean, median, mode, and standard deviation can provide insight into the characteristics of the data.

Conclusion

Ensuring data quality is one of the most pressing challenges today for most organizations. With applications constantly receiving new data and undergoing incremental changes, achieving data quality cannot be a onetime event. As organizations' appetite for big data grows daily in their quest to satisfy customers, suppliers, investors, and employees, the common obstacle of impediment is data quality. Improving data quality is the lynchpin to a better enterprise, better decision-making, and better functionality.

Data quality can be improved, and there are methods for doing so that are rooted in logic and experience. On the market are commercial off-the-shelf (COTS) products which are simple, intuitive methods to manage and analyze data – and establish business rules for an enterprise. Some can implement a data quality layer that filters any number of sources for quality standards; provide real-time monitoring; and enable the profiling of data prior to absorption and aggregation with a company's core data. At times, however, it will be necessary to bring in objective, third-party subject-matter experts for an impartial analysis and solution of an enterprise-wide data problem.

Whatever path is chosen, it is important for an organization to have a master data management (MDM) plan no differently than it might have a recruiting plan or a business development plan. A sound MDM creates an ever-present return

on investment (ROI) that saves time, reduces operating costs, and satisfies both clients and stakeholders.

Further Reading

- Capehart, B. L., & Capehart, L. C. (2005). Web based energy information and control systems: case studies and applications, 436–437.
- Cappiello, C., Francalanci, C., & Pernici, B. (2003). Time-related factors of data quality in multi-channel information systems. *Journal of Management Information Systems*, 20(3), 71–91.
- Friedman, T., & M. Smith. (2011). *Measuring the business value of data quality* (Gartner ID# G00218962). Available at: http://www.data.com/export/sites/data.com/mon/assets/pdf/DS_Gartner.pdf.
- Greengard, S. (1998). Don't let dirty data derail you. *Workforce*, 77(11), 107–108.
- Knolmayer, G., & R  thlin, M. (2006). Quality of material master data and its effect on the usefulness of distributed ERP systems. *Lecture Notes in Computer Science*, 4231, 362–371.
- Lederman, R., Shanks, G., Gibbs, M.R. (2003). Meeting privacy obligations: the implications for information systems development. Proceedings of the 11th European Conference on Information Systems. Paper presented at ECIS: Naples, Italy.
- Liebenau, J., & Backhouse, J. (1990). *Understanding information: an introduction*. Information systems. Palgrave Macmillan, London, UK.
- Loshin, D. (2006). *The data quality business case: Projecting return on investment* (White paper). Available at: http://knowledge-integrity.com/Assets/data_quality_business_case.pdf.
- Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. *Database Marketing & Customer Strategy Management*, 12(2), 105–112. Available at: <http://link.springer.com/article/10.1057/palgrave.dbm.3240247>.
- Redman, T. C. (1995). Improve data quality for competitive advantage. *MIT Sloan Management Rev.*, 36(2), pp. 99–109.
- Sellar, S. (1999). Dust off that data. *Sales and Marketing Management*, 151(5), 71–73.
- Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, 41(2), 54–57.
- Vayghan, J. A., Garfinkle, S. M., Walenta, C., Healy, D. C., & Valentin, Z. (2007). The internal information transformation of IBM. *IBM Systems Journal*, 46(4), 669–684.
- Watts, S., Shankaranarayanan, G., & Even, A. (2009). Data quality assessment in context: A cognitive perspective. *Decision Support Systems*, 48(1), 202–211.

Big Data R&D

► [Big Data Research and Development Initiative \(Federal, U.S.\)](#)

Big Data Research and Development Initiative (Federal, U.S.)

Fen Zhao¹ and Suzi Iacono²

¹Alpha Edison, Los Angeles, CA, USA

²OIA, National Science Foundation, Alexandria, VA, USA

Synonyms

[BD hubs](#); [BD spokes](#); [Big data hubs and spokes](#); [Big data R&D](#); [Data science](#); [Harnessing the data revolution](#); [HDR](#); [NSF](#)

Introduction

On March 29, 2012, the Office of Science and Technology Policy (OSTP) and the Committee on Technology Networking and Information Technology Research and Development Subcommittee (NITRD) launched the Federal Big Data Research and Development (R&D) Initiative. Since then, the National Science Foundation has been a leader across federal agencies in supporting and catalyzing Big Data research, development, and innovation across the scientific, public, and private sectors. This entry summarizes the timeline of Big Data and data science activities initiated by the NSF since the start of the initiative.

The Fourth Paradigm

Over the course of history, advances in science and engineering have depended on the development of new research infrastructure. The advent of

microscopes, telescopes, undersea vehicles, sensor networks, particle accelerators, and scientific observatories has opened windows into our observation and understanding of natural, engineered, and social phenomenon. For scientists, access to these instruments unlocks a myriad of new theories and approaches to the discovery process to help them develop new and different kinds of insights and breakthroughs. Many of these results have big payoffs for society – helping find novel solutions to challenges in health, education, national security, disaster prevention and mitigation, the economy, and the scientific discovery process itself.

Today, most research instruments produce very large amounts of data. Extracting knowledge from these increasingly large and diverse datasets requires a transformation in the culture and conduct of scientific discovery. Some have referred to this movement as the *Fourth Paradigm* (Tansley and Tolle 2009), where the unique capabilities of data science have defined a new mode of scientific discovery and a new era of scientific progress. Take, as an illustrative example, the revolution occurring in oceanography. Oceanographers are no longer limited, as they had been in previous decades, to the small amounts of data they can collect in summer research voyages. Now, they can remotely collect data from big sensor networks at the bottom of the sea or in estuaries and rivers and conduct real-time analysis of that data. This story of innovation is echoed in countless scientific disciplines, as having more access to complete datasets and advanced analytic techniques spurs faster insights and improved hypotheses.

Harnessing this so-called data revolution has enormous potential impact. That is why the National Science Foundation (NSF) recently announced that *Harnessing Data for 21st Century Science and Engineering* (HDR) would be one of *NSF's 10 Big Ideas*. The Big Ideas are a set of bold ideas for the Foundation that look ahead to the most impactful trends in the future of science and society:

... NSF proposes *Harnessing Data for 21st Century Science and Engineering*, a bold initiative to

develop a cohesive, national- scale approach to research data infrastructure and a 21st-century workforce capable of working effectively with data.

This initiative will support basic research in math, statistics and computer science that will enable data-driven discovery through visualization, better data mining, machine learning and more. It will support an open cyberinfrastructure for researchers and develop innovative educational pathways to train the next generation of data scientists.

This initiative builds on NSF's history of data science investments. As the only federal agency supporting all fields of S&E, NSF is uniquely positioned to help ensure that our country's future is one enriched and improved by data.

What's All the Excitement About?

By 2017, the term “Big Data” has become a common buzzword in both the academic and commercial sectors. International Data Corporation (IDC) regularly makes predictions about the growth of data: “The total amount of data in the world was 4.4 zettabytes in 2013. That is set to rise steeply to 44 zettabytes by 2020. To put that in perspective, one zettabyte is equivalent to 44 *trillion* gigabytes. This sharp rise in data will be driven by rapidly growing daily production of data” (Turner et al. 2014). Now, IDC believes that by 2025 the total will hit 180 zettabytes.

Complementary to the continuing focus on Big Data, many thought leaders today have started to emphasize the importance of “little data” or data at the long tail of science. There are thousands of scientists who rely on their own methods to collect and store data at small or medium scales in their offices and labs. The ubiquity and importance of data at all scales within scientific research has led NSF to mandate that every proposed project includes a data management plan. This data management plan is a critical part of the merit review of that project, regardless of the size of the datasets they are collecting and analyzing.

The importance of data at the long tail of science illustrates that some of the most exciting facets of data innovation do not center only on scale. Often experts will cite a framework of the

so-called Four Vs of Big Data, which help summarize the challenges and opportunities for data: *volume*, *variety*, *velocity*, and *value*. The ability to integrate a *variety* of types of data across different areas of science allows us to target grand challenges whose solution continues to elude us when approached from a single discipline viewpoint – for example, leveraging data between or among fields observing the earth, ocean, and/or atmosphere can help us to answer the biggest and most challenging research questions about our environment as a whole. Similarly, our ability to handle data at the *velocity* of life is necessary for addressing the many challenges that must be acted upon in real time – for example, responding to storms and other disasters. Yet, technology imposes many limitations on updating simulations and models and supporting decision-making on the ground in the pressing moment of need. Finally, understanding and quantifying the *value* of the massive numbers of diverse datasets now collected by all sectors of society is still an open question for the data science research community. Understanding the value and use of datasets is critical to finding solutions to major challenges around curation, reproducibility, storage, and long-term data management as well as to privacy and security considerations. If our research communities can resolve today's many hard problems around data science, benefits will be global while strengthening opportunities for US leadership.

Kicking Off a Federal Big Data Research and Development Initiative

In December 2010, the President's Council of Advisors for Science and Technology (PCAST) report, *Designing a Digital Future: Federally Funded Research Development in Networking and Information Technology* (Holdren 2010), challenged the federal research agencies to take actions to support more research and development (R&D) on Big Data.

Shortly after that, the Office of Science and Technology Policy (OSTP) responded to

this call and chartered an interagency Big Data Senior Steering Group (BDSSG) under the Committee on Technology Networking and Information Technology Research and Development Subcommittee (NITRD). NSF and the National Institutes of Health (NIH) have co-chaired this group over the years, while approximately 18 other research agencies have sent representatives to the meetings. Several non-research federal agencies with interests in data technologies also participated informally in the BDSSG.

Over the course of the years following its establishment, the BDSSG inventoried the existing Big Data programs and projects at each of the participating agencies and began coordinating across the agencies. Efforts were divided into four main areas: investments in Big Data foundational research, development of cyber-infrastructure in support of domain-specific data-intensive science and engineering, support for data science education and workforce development, and activities in support of increased collaboration and partnerships with the private sector. Other important areas, such as privacy and open access, were also identified as critical to Big Data by the group but became the focus areas of new NITRD subgroups – for example, the Privacy Research and Development Interagency Working Group (Privacy R&D IWG). Big Data R&D remained the central focus of the work of the BDSSG.

On March 29, 2012, OSTP and NITRD launched the Federal Big Data Research and Development Initiative across federal agencies. The NSF press release states:

At an event led by the White House Office of Science and Technology Policy in Washington, D.C., (then NSF Director) Suresh joined other federal science agency leaders to discuss cross-agency big data plans and announce new areas of research funding across disciplines in this field.

NSF announced new awards under its Cyber-infrastructure for the 21st Century framework and Expeditions in Computing programs, as well as awards that expand statistical approaches to address big data. The agency is also seeking proposals under a Big Data solicitation, in collaboration

with the National Institutes of Health (NIH), and anticipates opportunities for cross-disciplinary efforts under its Integrative Graduate Education and Research Traineeship program and an Ideas Lab for researchers in using large datasets to enhance the effectiveness of teaching and learning.

About 11 other agencies also participated. The White House Big Data Fact Sheet included their announcements. Here are some examples:

- DARPA launched the XDATA program, which sought to develop computational techniques and software tools for analyzing large volumes of semi-structured and unstructured data. Central challenges to be addressed included scalable algorithms for processing imperfect data in distributed data stores and effective human-computer interaction tools that are rapidly customizable to facilitate visual reasoning for diverse missions.
- DHS announced the Center of Excellence on Visualization and Data Analytics (CVADA), a collaboration among researchers at Rutgers University and Purdue University (with three additional partner universities each) that lead research efforts on large, heterogeneous data that First Responders could use to address issues ranging from man-made or natural disasters to terrorist incidents, law enforcement to border security concerns, and explosives to cyber threats.
- NIH highlighted The Cancer Genome Atlas (TCGA) project – a comprehensive and coordinated effort to accelerate understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.

Collectively, these activities had an impressive impact; in a 2013 report, PCAST commended the agencies for their efforts to push Big Data into the forefront of their research priorities: “Federal agencies have made significant progress in supporting R&D for data collection, storage, management, and automated large-scale analysis (Holdren 2013). They recommended continued emphasis on these investments in future years.

Taking the Next Steps: Developing National, Multi-stakeholder Big Data Partnerships

Entering the second year of the Big Data Initiative, the BDSSG expanded the framing of the Federal Big Data R&D Initiative as a coordinated national endeavor rather than just a federal government effort. To encourage the participation of stakeholders in private industry, academia, state and local government, nonprofits, and foundations to develop and participate in Big Data Initiatives across the country, in April 2013, NSF issued a request for information (RFI) about Big Data. This RFI encouraged non-federal stakeholders to identify the kinds of Big Data projects they were willing to participate in to further Big Data innovation across the country. Of particular interest were cross sector partnerships designed to advance core Big Data technologies, harness the power of Big Data to advance national goals, initiate new competitions and challenges, and foster regional innovation.

In November 2013, the BDSSG and NITRD convened *Data to Knowledge to Action* (Data 2Action), an event in Washington, DC, that highlighted a number of high-impact, novel, multi-stakeholder partnerships surfaced through the RFI and later outreach efforts. These projects embraced collaboration between the public and private sectors and promoted the sharing of data resources and the use of new sophisticated tools to plumb the depths of huge datasets and derive greater value for American consumers while growing the nation's economy.

The event featured scores of announcements by corporations, educational institutions, professional organizations, and others that – in collaboration with federal departments and agencies and state and local governments – enhance national priorities such as economic growth and job creation, education and health, energy and sustainability, public safety and national security, and global development. About 30 new partnerships were announced with a total of about 90 partners.

Some examples include:

- A new Big Data analytics platform Spark created by UC Berkeley's AMPLab which was funded by NSF, DARPA, DOE, and a host of private companies such as Google and SAP.
- A summer program called Data Science for Social Good (funded by the Schmidt Family Foundation and University of Chicago with partners including the City of Chicago, Cook County Land Bank, Cook County Sheriff, Lawrence Berkeley National Labs, and many others) hosted fellows to create applications to solve data science challenges as defined by their partners.
- Global corporations Novartis, Pfizer, and Eli Lilly partnered to improve access to information about clinical trials, including matching individual health profiles to applicable clinical trials.

While the Data 2Action event catalyzed federal outreach on Big Data research to communities beyond academia, continuing Big Data innovation on a national scale required sustained community investment beyond federal coordination. To help achieve this goal of sustained community dialogue and partnerships around Big Data, in the fall of 2014, NSF's Directorate for Computer and Information Science and Engineering (CISE) announced a plan to establish a National Network of *Big Data Regional Innovation Hubs* (BD Hubs). Released in winter 2015, the program solicitation to create the Hubs states:

Each BD Hub [is] a consortium of members from academia, industry, and/or government... [and is] across distinct geographic regions of the United States, including the Northeast, Midwest, South, and West... and focus[es] on key Big Data challenges and opportunities for its region of service. The BD Hubs aim to support the breadth of interested local stakeholders within their respective regions, while members of a BD Hub should strive to achieve common Big Data goals that would not be possible for the independent members to achieve alone.

To foster collaboration among prospective partners within a region, in April 2015, NSF sponsored a series of intensive 1-day "charrettes" to convene

stakeholders, explore Big Data challenges, and aid in the establishment of the Hub consortia. “Charrettes” are meeting in which all stakeholders in a project attempt to resolve conflicts and map solutions. NSF convened a charrette in each of the four Hub geographic regions. To facilitate discussion beyond the charrette, a HUBzero community portal was established over the course of the initial Hub design and implementation process. Potential partners used this portal to communicate with other members or potential partners within their Hub.

In November 2015, NSF announced seven awards totaling more than \$5 million to establish four regional Hubs for data science innovation. The consortia are coordinated by top data scientists at Columbia University (Northeast Hub), Georgia Institute of Technology with the University of North Carolina (South Hub), University of Illinois at Urbana-Champaign (Midwest Hub), and University of California, San Diego, University of California, Berkeley, and University of Washington (West Hub).

Covering all 50 states, they include initial partnership commitments from more than 250 organizations. These organizations ranged from universities and cities to foundations and Fortune 500 corporations, and the four Hubs developed plans to expand the consortia further over time. The network of four Hubs established a “big data brain trust” geared toward conceiving, planning, and supporting Big Data partnerships and activities to address regional challenges. Among the benefits of the program for Hub members are greater ease in initiating partnerships by reducing coordination costs; opportunities for sharing ideas, resources, and best practices; and access to top data science talent.

While Hubs focused primarily on ideation and coordination of regional Big Data partnerships, additional modes of support were needed for the actual projects that were to become the outputs of those coordination efforts. These projects were called the “Spokes” of the Big Data Hub network. The Spokes were meant to focus on data innovation in specific areas of interest, for example, drought data archives in the west or health data

on underrepresented minorities in the south. In Fall 2015, NSF solicited proposals for Spokes projects that would work in concert with their corresponding regional BD Hub to address one of three broad challenges in Big Data:

- Accelerating progress towards addressing societal grand challenges relevant to regional and national priority areas;
- Helping automate the Big Data lifecycle; and
- Enabling access to and increasing use of important and valuable available data assets, also including international datasets....

Similar to a Hub, each Big Data Spoke takes on a convening and coordinating role as opposed to conducting fundamental research. Unlike a Hub, each Spoke would have a specific goal-driven scope within an application or technology area. Typical Spoke activities included, for example, gathering important stakeholders via forums, meetings, or workshops; engaging with end users and solution providers via competitions and community challenges; and forming multidisciplinary teams to tackle questions no single field could solve alone. Strategic leadership guiding both the Big Data Hubs and Spokes comes from each Hub’s Steering Committee – a group of Big Data experts and thought leaders across sectors that act as advisors and provide general guidance.

In 2016 and 2017, NSF awarded \$13 million to 11 Spoke projects, 10 planning grants, and a number of other Spoke-related projects. Project topics range from precision agriculture to personalized education and from data sharing to reproducibility. The range of Spoke topics reflected the unique priorities and capabilities of the four Big Data Hubs and their regional interests. A second Spoke solicitation was released in March 2017, and new awards are expected by the end of fiscal year 2018.

Developing an Interagency Strategic Plan

Starting in 2014, the BDSSG began work on an interagency strategic plan to help coordinate

future investments in Big Data R&D across the federal research agencies. A key assumption of this plan was that it would not be prescriptive at any level, but instead would be a potential enabler of future agency actions by surfacing areas of commonalities and priority to support agency missions. The development of this strategic plan was supported through a number of cross-agency workshops, a request for information from the public, and a workshop with non-federal stakeholders to gauge their opinions.

Building upon all the work that had been carried out to date on the National Big Data R&D Initiative, *the Federal Big Data Research and Development Strategic Plan* (Plan) [NITRD

2016] aimed to “build upon the promise and excitement of the myriad applications enabled by Big Data with the objective of guiding Federal agencies as they develop and expand their individual mission-driven programs and investments related to Big Data.” The Plan described a vision for Big Data innovation shared across federal research agencies (Fig. 1):

We envision a Big Data innovation ecosystem in which the ability to analyze, extract information from, and make decisions and discoveries based upon large, diverse, and real-time datasets enables new capabilities for Federal agencies and the Nation at large; accelerates the process of scientific discovery and innovation; leads to new fields of research and new areas of inquiry that would otherwise be

Big Data Research and Development Initiative (Federal, U.S.),

Fig. 1 The cover of the Federal Big Data Research and Development Strategic Plan (2016)



impossible; educates the next generation of 21st century scientists and engineers; and promotes new economic growth.

The Plan articulates seven strategies that represent key areas of importance for US Big Data R&D. These are:

- Strategy 1: Create next-generation capabilities by leveraging emerging Big Data foundations, techniques, and technologies.
- Strategy 2: Support R&D to explore and understand trustworthiness of data and resulting knowledge, to make better decisions, enable breakthrough discoveries, and take confident action.
- Strategy 3: Build and enhance research cyber-infrastructure that enables Big Data innovation in support of agency missions.
- Strategy 4: Increase the value of data through policies that promote sharing and management of data.
- Strategy 5: Understand Big Data collection, sharing, and use with regard to privacy, security, and ethics.
- Strategy 6: Improve the national landscape for Big Data education and training to fulfill increasing demand for both deep analytical talent and analytical capacity for the broader workforce.
- Strategy 7: Create and enhance connections in the national Big Data innovation ecosystem.

While the strategic plan addresses the challenges outlined by Four Vs of Big Data, it encompasses a broader vision for the future of data science and its application toward mission and national goals. Strategy 2 emphasizes the need to move beyond managing scale to enabling better use of data analytics outputs; rather than focusing on the first part of the “Data to Knowledge to Action” pipeline, which is usually focused on purely technological solutions, it recognizes the need to understand the sociotechnical needs to derive actionable insight from data-driven knowledge.

Strategies 3 and 4 both address the national need to sustain an ecosystem of open data and the tools to analyze that data; such an infrastructure supports not only federal agency missions but the utility of Big Data to the private sector and the public at large. Agencies also recognized the risks and challenges in using Big Data in developing Strategy 5, focusing not only on the privacy, security, and ethical challenges that come with Big Data analytics today but pressing for more

research on how to reduce risk and maximize benefits for the data-driven technologies of the future.

Multiple industry reports (Manyika et al. 2011) have forewarned of a dramatic and continuing deficit for demand in data analytics talent within the USA. This deficit ranges from data-savvy knowledge workers to PhD-trained data scientists. Research agencies acknowledge the needs for programs that support the development of a workshop with data skills at all levels to staff, support, and communicate their mission programs.

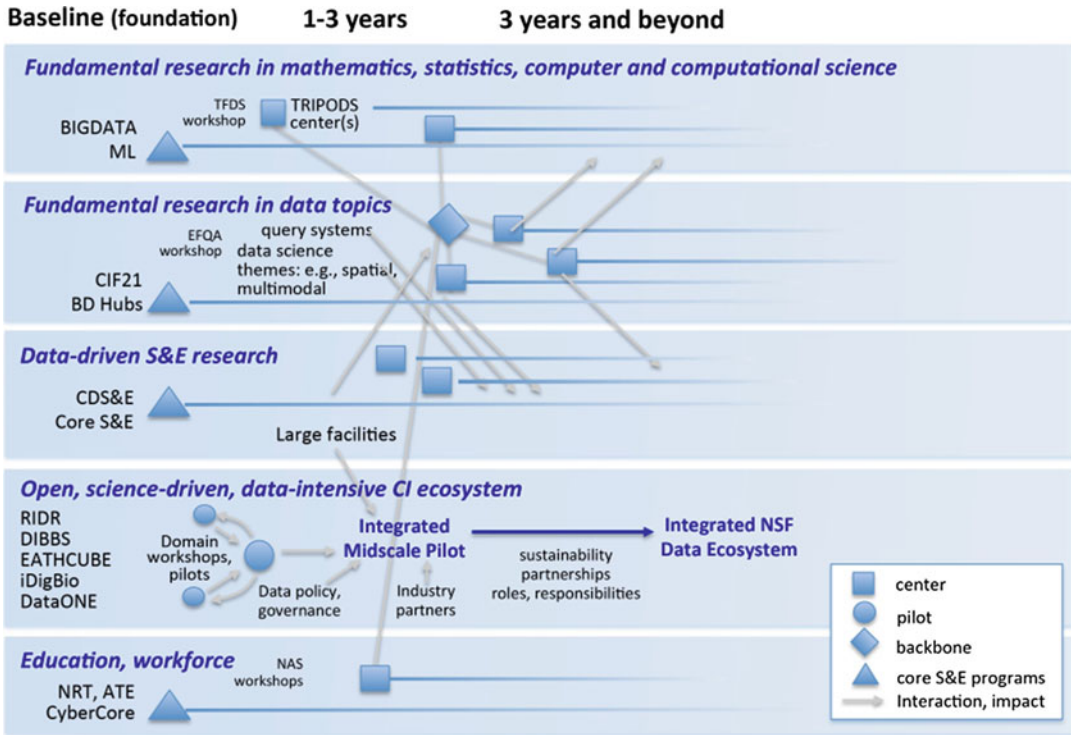
Through the strategic planning process, agencies saw many synergies between different agency missions in their use of Big Data. Strategy 7 acknowledges ways that agencies could act collectively to create interagency programs and infrastructures to share benefits across the federal government.

Moving Toward the Future: Harnessing the Data Revolution

NSF’s Harnessing the Data Revolution (HDR) Big Idea builds on past investments and lays the foundations for the future transformation of science and society by data technologies. HDR has a number of major themes, which are outlined in Fig. 2.

Given NSF’s breadth of influence over almost all fields of science, the Foundation can bring together science, engineering, and education experts into convergence teams to make this vision a reality. NSF has a unique role within universities (which are critical participants) in the support of research, sustainable research infrastructure, and development of human capital. NSF also has strong connections with industry and with funding agencies around the world. Given the trend toward global science and the value of sharing research data internationally, NSF is well positioned to work with other research agencies when moving forward on research priorities.

HDR’s thesis on foundational theory-based research in data science is that it must exist at the intersection of math, statistics, and computer



Big Data Research and Development Initiative (Federal, U.S.), Fig. 2 Conceptualization of NSF's Harnessing the Data Revolution (HDR) Big Idea

science. Experts in each of these three disciplines must leverage the unique perspective of their field in unison to develop the next generation of methods for data science. The TRIPODS program recognizes this needed convergence by funding data science center-lets (pre-center scale grants) that host experts across all three disciplines.

Today, research into the algorithms and systems for data science must be sociotechnical in nature. New data tools must not only manage for the Four Vs but also the challenges of human error or misuse. New systems are needed to help data user understand the limits of their statistical analysis, manage privacy and ethical issues, ensure reproducibility of their results, and efficiently share data with others.

Promoting progress in the scientific disciplines is the core of NSF's mission. At the heart of the HDR Big Idea is the tantalizing potential of leap-frogging progress in multiple sciences through applications of translational data science. The

benefits of advanced data analytics and management could be leveraged by all size of science research projects, from individual to center scale.

One of the key components of the HDR Big Idea is the design and construction of a national data infrastructure of use to a wide array of science and engineering communities supported by NSF. Past investments by NSF have built some basic components of this endeavor, but others have yet to be imagined. The ultimate goal is a co-designed, robust, comprehensive, open, science-driven, research cyber-infrastructure (CI) ecosystem capable of accelerating a broad spectrum of data-intensive research, including research in large scale and Major Research Equipment and Facilities Construction (MREFC).

Innovative learning opportunities and educational pathways are needed to build a twenty-first-century data-capable workforce. These opportunities must be grounded in an

education research-based understanding of the knowledge and skill demands needed by that workforce. NSF's current education programs span the range from informal, K-12, undergraduate, graduate, and postgraduate education and can be leveraged to train the full diversity of this nation's workforce needs. Of particular interest is undergraduate and graduate data science training, to create pi-shaped scientists who have broad skills but with deep expertise in data science in addition to their scientific domain.

In Summary

This short entry summarizes some of the work done for the National Big Data Research and Development Initiative since the beginning of calendar year 2011. It is written from the point of view of NSF because that is where the authors are located. It should be noted one could imagine different narratives if it were written from the NIH or DARPA perspectives with stories that would be equally compelling.

Through a reorganization of the NITRD Subcommittee, the BDSSG was renamed the Big Data Interagency Working Group (BDIWG), showing its intention to be a permanent part of that organization. The current co-chairs are Chaitan Baru from NSF and Susan Gregurick from NIH.

But the work continues.

Further Reading

- Holdren, J. P., Lander, E., & Varmus, H. (2010). Report to the president and congress: Designing a digital future: Federally funded research and development in networking and information technology. *Executive Office of the President and President's Council of Advisors on Science and Technology*. <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf>.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey & Company. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx.

- Tansley, S., & Tolle, K. M. (2009). In T. Hey (Ed.), *The fourth paradigm: Data-intensive scientific discovery* (Vol. 1). Redmond: Microsoft Research.
- Turner, V., Gantz, J. F., Reinsel, D., & Minton, S. (2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. *IDC Analyze the Future*. https://scholar.google.com/scholar?cluster=2558441206898490167&hl=en&as_sdt=2005&sciodt=0,5.

Big Data Theory

Melanie Swan

New School University, New York, NY, USA

Definition/Introduction

Big Data Theory is a set of generalized principles that explain the foundations, knowledge, and methods used in the practice of data-driven science.

Part I: Theory of Big Data in General

In general a theory is an explanatory mechanism. A theory is a supposition or a system of ideas intended to explain something, especially one based on general principles independent of the thing to be explained. Big Data Theory explains big data (data-driven science), what it is and its foundations, approaches, methods, tools, practices, and results. A theory explains something in a generalized way. A theory attempts to capture the core mechanism of a situation, behavior, or phenomenon.

A theory is a class of knowledge. Different classes of knowledge have different proof standards. The overall landscape of knowledge includes observation, conjecture, hypothesis, prediction, theory, law, and proof. Consider Newton's laws, for example, and the theory of gravity; laws have a more-established proof standard than theories. An explanation of a phenomenon is called a theory, whereas a law is a more formal description of an observed phenomenon.

Many theories do not become laws but serve as a useful tool for practitioners to understand and work with a phenomenon practically.

Here are some examples of theories in which the same structure would apply to theories of data-driven science (a theory provides an explanatory mechanism):

- Darwin's theory of evolution states that all species of organisms arise and develop through the natural selection of small, inherited variations that increase the individual's ability to compete, survive, and reproduce.
- Pythagoras's theorem is a fundamental relation in Euclidean geometry among the three sides of a right triangle, stating that the square of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the other two sides.
- Bayes theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event
- A theory of error estimation is a set of principles on which the practice of the activity of error estimation is based.

There is the Theory of Big Data and the Philosophy of Big Data. Theory relates to the internal practices of a field, and philosophy includes both the internal practices of the field and the external impact, the broader impact of the field on the individual and society. The Philosophy of Big Data is the branch of philosophy concerned with the definition, methods, and implications of big data and data-driven science in two domains:

- Internal Industry Practice: Internal to the field as a generalized articulation of the concepts, theory, and systems that comprise the overall use of big data and data science
- External Impact: External to the field, considering the impact of big data more broadly on individuals, society, and the world, for example, addressing data concerns such as security, privacy, and accessibility.

The Philosophy of Big Data and Data-driven Science may have three areas. These include

ontology (existence; the definition, dimensions, and limitations of big data), epistemology (knowledge; the knowledge obtained from big data and corresponding proof standards), and axiology (valorization; ethical practice, the parts of big data practices and results that are valorized as being correct, accurate, elegant, right).

The Philosophy of Big Data or Data Science is a branch of philosophy concerned with the foundations, methods, and implications of big data and data science. Big data science is scientific practices that extract knowledge from data using techniques and theories from mathematics, statistics, computing, and information technology. The philosophical concerns of the Philosophy of Big Data Science include the definitions, meaning, knowledge production possibilities, conceptualizations of science and discovery, definitions of knowledge, proof standards, and practices in situations of computationally intensive science involving large-scale, high-dimensional modeling, observation, and experimentation in network environment with very-large data sets.

Part II: Theory in Big Data and Data-Driven Science

The Theories of Big Data and data-driven science correspond to topic areas of big data and data-driven science. Instead of having a "data science theory" overall which might be too general a topic to have an explanatory theorem, theories are likely to relate to topics within the field of data science. For example, there are theories of machine learning, Bayesian updating, and classification and unstructured learning.

Big data, data science, or data-driven science is an interdisciplinary field using scientific methods, processes, and systems to extract knowledge and insights from various forms of data. The primary method of data science is statistics; however, other mathematical methods are also involved such as linear algebra and calculus. In addition to mathematics, data science may involve information theory, computation, visualization, and other methods of data collection, modeling, analysis, and decision-making. The term "data science"

was used by Peter Naur in 1960 as a synonym for computer science.

Theories of data science may relate to the kinds of activities in the field such as description, prediction, evaluation, data gathering, results communication, and education. Theories may correspond to the kinds of concepts used in the field such as causality, validity, inference, and deduction. Theories may address foundational concerns of the field, for example, general principles (the bigger the data corpus, the better), and commonly used practices (how p-values are used). Theories may be related to methods within an area, for example, a theory of structured or unstructured learning. There may be theories of shallow learning (1–2 layers) relating to methods specific to that topical area such as Bayesian inference, support vector machines, decision trees, K-means clustering, and K-nearest neighbor analysis. Similarly, there may be theories of deep learning (5–20 layers) relating to methods of practice specific to that area such as neural nets, convolutional neural nets in the case of image recognition, and recurrent neural nets in the case of text and speech recognition.

Specific data science topics are the focus in Big Data Theory workshops to address situations where a theory with an explanatory mechanism would be useful in a generalized sense beyond specific use cases. Some of these topics include:

- Practices related to model-fit, “map-territory,” explanandum-explanans (fit of explanatory model to that which is to be explained), scale-free models, and model-free learning
- Challenges of working with big data problems such as spatial and temporal analysis, time series, motion data, topological data analysis, hypothesis-testing, computational limitations and cloud computing, distributed and network-based computing models, graph theory, complex applications, results visualization, and big data visual analytics
- Concepts such as randomization, entropy, evaluation, adaptive estimation and prediction, error, multivariate volatility, high-dimensional operations (causal inference, estimation, learning), and hierarchical modeling
- Mathematical methods such as matrix operations, optimization, least-squares, gradient descent, structural optimization, Bayesian updating, regression analysis, linear transformation, scale inference, variable clustering, model-free learning, pattern analysis, and kernel methods
- Industry norms related to data-sharing and collaboration models, peer review, and experimental results replication

Conclusion

Big Data Theory is a set of generalized principles that explain the foundations, knowledge, and methods used in the practice of data-driven science. The Philosophy of Big Data or Data Science is a branch of philosophy concerned with the foundations, methods, and implications of big data and data science. Big data science is scientific practices that extract knowledge from data using techniques and theories from mathematics, statistics, computing, and information technology. The philosophical concerns of the Philosophy of Big Data Science include the definitions, meaning, knowledge production possibilities, conceptualizations of science and discovery, definitions of knowledge, proof standards, and practices in situations of computationally intensive science involving large-scale, high-dimensional modeling, observation, and experimentation in network environment with very-large data sets.

Further Reading

- Harris, J. (2013). The need for data Philosophers. *The Obsessive-Compulsive Data Quality (OCDQ) blog*. Available online at <http://www.ocdqblog.com/home/the-need-for-data-philosophers.html>.
- Swan, M. (2015). Philosophy of Big Data: Expanding the human-data relation with Big Data science services. *IEEE BigDataService 2015*. Available online at http://www.melanieswan.com/documents/Philosophy_of_Big_Data_SWAN.pdf.
- Symons, J., & Alvarado, R. (2016). Can we trust Big Data? Applying philosophy of science to software. *Big Data & Society*, 3(2), 1–17. Available online at <http://journals.sagepub.com/doi/abs/10.1177/2053951716664747>.

Big Data Workforce

Connie L. McNeely and Laurie A. Schintler
George Mason University, Fairfax, VA, USA

The growth of big data, its engagement, and its applications are growing within and across all sectors of society and, as such, require a big data workforce. Big data encompasses processes and technologies that can be applied across a wide range of domains from business to science, from government to the arts (Economist 2010). Understanding this workforce situation calls for examination from not only technical but, importantly, social and organizational perspectives. Also, the skills and training necessary for big data related jobs in industry, government, and academia have become a focus of discussions on educational attainment relative to workforce trajectories.

The complex and rapidly changing digital environment is marked by the growth and spread of big data and by related technologies and activities that pose workforce development challenges and opportunities that require specific and evolving skills and training for the changing jobs landscape. Big data and calls for related workers appear across virtually all sectors (Galov 2020). The range of areas in which big data plays an increasingly central role requires agile and flexible workers with the ability to rapidly analyze massive datasets derived from multiple sources in order to provide information to enable actions in real-time. As one example, big data may enable more precise dosing of medications and has been used to develop sensor technologies to determine when a football player needs to be side-lined due to heightened risks of concussion (Frehill 2015, p. 58). Yet another example is high-frequency trading, which draws upon various sources of real-time data to create real-time actionable insights.

Understanding the relationship between the workforce needs of big data employers and the supply of workers with skills adapted to related positions is a key consideration in determining what constitutes the big data workforce. Although posing challenges to strict labor classification, its

principal characterizing elements can be distinguished according to 1) skill and task identification, 2) disciplinary and field delineations, and 3) organizational specifications.

Skills-Based Classification

Engaging and applying data in work processes has become a basic requirement in many jobs and has led to new job creation in a number of areas. Big data are calling for more and more workers with “deep skills and talent,” pointing to the relationship between higher education and the development of the big data workforce. However, organizational needs in this regard are variable. For example, data analytics are now fundamental to positions such as management analysts and market research analysts, both of which require only short-term certification, as opposed to longer degree terms of study (Carantit 2018). Some of the basic technical skills required to handle big data are accessing databases to query data, ingesting data into databases, gathering data from various sources using web scraping, parsing, and tokenizing texts in big data storage environments (NASEM 2018).

Technically speaking, training and education for many big data jobs typically require a basic knowledge of statistics, quantitative methods, or programming, upon which applicable skillsets can be built. More than the computing sciences, such background can be acquired in a number of fields that have long incorporated related preparation. As one example, “social scientists have worked with exceptionally large datasets for quite some time, historically accessing remote space, writing code, analyzing data, and then telling stories about human social behavior from these complex sources.” Indeed, many “techniques, tools, and protocols developed by social science research communities to manage and share large datasets – including attention to the ethical issues associated with collecting these data – hold important implications for the big data workforce” (Frehill 2015, pp. 49, 52).

General descriptions have indicated that to exploit big data – characterized particularly by

velocity, variety, and volume – workers are needed with “the skills of a software programmer, statistician, and storyteller/artist to extract nuggets of gold hidden under mountains of data” (Economist 2010). These characteristics are encompassed in the broad occupational category of “data scientist” (Hammerbacher 2009). Considering the combination of disparate skills required to capture value from big data, three key types of workers have been identified under the rubric of data scientist (Manyika et al. 2011):

1. *Deep analytical talent* – people with technical skills in statistics and machine learning, for example, capable of analyzing large volumes of data to derive business insights.
2. *Data-savvy managers and analysts* who have the skills to be effective consumers of big data insights, that is, capable of posing the right questions for analysis, interpreting and challenging the results, and making appropriate decisions.
3. *Supporting technology personnel* who develop, implement, and maintain the hardware and software tools, such as databases and analytic programs, needed to make use of big data.

Note that such skills and workers – deep analytical talent, data-savvy managers and analysts, and supporting technology personnel – principally apply to capacities and capabilities to extract information from massive amounts of data and to enabling related data-driven decision-making in work settings. However, disciplinary silos complicate the picture of the big data workforce and associated occupational needs. These skills are required in various fields. The arena from which data scientists are drawn and in which associated skills are developed is broader than the pool of those trained in computing and information technology disciplines, with many being basic requirements in various liberal arts fields, including social sciences and other science and technology areas, ranging from, for example, architects to sociologists to engineers. Moreover, technology, sources, and applications of big data, big data analytics, big data hardware, and big data storage

and processing capacities are constantly evolving. As such, the skillset for the big data worker is something of a moving target, and depending on how skill requirements are specified, the type and size of the pool of big data workforce talent can vary accordingly (Frehill 2015).

Skill Mismatch Dilemmas

Frankly, relative to employer practices and workforce needs, when an industry or field is growing rapidly, “it is not unusual for a shortage of workers to occur until educational institutions and training organizations build the capacity to teach more individuals, and more people are attracted to the needed occupations” (CEA 2014, p. 41), a point that is translated in the growing number of analytics and data science programs (Topi and Markus 2015). However, rapidly accelerating big data growth and technological change can pose limits to skills forecasting. Accordingly, some recommendations focus on gaining adaptable core, transversal skills and on building technical learning capacities, rather than on planning education and training to meet specified forecasts of requirements, especially since they may change before curricular programs can adjust. “Shorter training courses, which build on solid general technical and core skills, can minimize time lags between the emergence of skill needs and the provision of appropriate training” (ILO 2011, p. 22). Be that as it may, especially given assertions of a skill mismatch and gap for manipulating, analyzing, and understanding big data, the relationship between education and the development of the big data workforce is a critical point of departure for delineating the field in general.

Skill mismatch, as a term, can relate to many forms of labor market friction and imbalance, including educational vertical and horizontal mismatches, skill gaps, skill shortages, and skill obsolescence (McGuinness et al. 2017). In general, skill mismatch refers to labor market imbalances and workforce situations in which skill supply and skill demand diverge. Such is the case with big data analytics and digital skill requirements relative to employer asserted shortages and needs.

Workforce Participation and Access

The rapid and dramatic changes brought about by big data in today's increasingly digitized society have led to challenges and opportunities impacting the related workforce. Education and training, hiring, and career patterns point to social and labor market conditions that reflect changes in workforce participation and representation. The ubiquitous nature of big data has meant an expanded need for workers with a variety of applicable skills (many of which entail relatively good earning potential). Against this backdrop, important questions have been raised about those who use it and those who work with it. Big data and related technologies and activities can mean increased demands and wages for highly skilled workers and, arguably, will hold more possibilities for employment opportunities and participation.

However, especially in light of socio-cultural and structural dynamics relative to labor market processes that shape and are subsequently shaped by demographic factors such as race, ethnicity, gender, disability, etc., questions of worker identity and skills are brought to the big data agenda, with particular attention to disparities in terms of educational and workforce dynamics. For example, minorities and women constitute only a small percentage of the big data workforce, signaling attention to capacity building and to questions of big data skill attainment and of workforce opportunity, access, participation, and mobility. Also, the use of big data in human resource activities affects recruitment and retention practices, with specific algorithms developed to monitor trends and gauge employee potential, in addition to general performance tracking and surveillance (Carantit 2018). Keeping in mind that a variety of social, political, and economic factors affect educational and skill attainment in the first place, such issues involve attention to the allocation of occupational roles, upgrading of skills, and access to employment opportunities. Big data and related digital skill requirements leave some individuals and groups at higher risk of unemployment and wage depression (e.g., women, minorities, and older, lower-

educated, and low-skill workers). All in all, the role of big data in shaping social, political, and economic relations (and power) come into play as reflected in educational and workforce opportunity and access.

Conclusion

New opportunities and prospects, but also new challenges, controversies, and vulnerabilities, have marked the explosion of big data and, so too, the workforce associated with it. Indeed, there is a need for big data workers "who are sensitive to data downsides as well as upsides" to achieve the benefits of big data while avoiding harmful consequences (Topi and Markus 2015, p. 39). The use of big data, along with machine learning and AI, is transforming economies and, arguably, delivering new waves of productivity (Catlin et al. 2015). Accordingly, educating, training, and facilitating access to workers with big data analytical skills is the sine qua non of the future.

Further Reading

- Berman, J. J. (2013). *Principles of big data: Preparing, sharing, and analyzing complex information*. Burlington: Morgan Kaufman.
- Carantit, L. (2018). Six ways big data has changed the workforce. <https://ihrim.org/2018/06/six-ways-big-data-has-changed-the-workforce>.
- Catlin, T., Scanlan, J., & Willmott, P. (2015, June 1). Raising your digital quotient. McKinsey Quarterly. <https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/raising-your-digital-quotient>.
- Chmura Economics and Analytics. (CEA). (2014). Big Data and Analytics in Northern Virginia and the Potomac Region. Northern Virginia Technology Council. <https://gwtoday.gwu.edu/sites/gwtoday.gwu.edu/files/downloads/BigData%20report%202014%20for%20Web.pdf>.
- Economist. (2010). Data, data everywhere. <http://www.economist.com/node/15557443>.
- Frehill, L. M. (2015). Everything old is new again: The big data workforce. *Journal of the Washington Academy of Sciences*, 101(3), 49–62.
- Galov, N. (2020, November 24). 77+ Big data stats for the big future ahead | updated 2020. <https://hostingtribunal.com/blog/big-data-stats>.

- Hammerbacher, J. (2009). Information platforms and the rise of the data scientist. In T. Segaran & J. Hammerbacher (Eds.), *Beautiful data: The stories behind elegant data solutions* (pp. 73–84). Sebastapol: O'Reilly.
- International Labour Organization (ILO). (2011). *A skilled workforce for strong, sustainable, and balanced growth: A G20 training strategy*. Geneva: International Labour Organization.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation>.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.
- McGuinness, S., Pouliakas, K., & Redmond, P. (2017). *How useful is the concept of skills mismatch?* Geneva: International Labour Organization.
- McNeely, C. L. (2015). Workforce issues and big data analytics. *Journal of the Washington Academy of Sciences*, 101(3), 1–11.
- National Academies of Sciences, Engineering, and Medicine (NASEM). (2018). *Data science for undergraduates: Opportunities and options*. Washington, DC: National Academies Press.
- Topi, H., & Markus, M. L. (2015). Educating data scientists in the broader implications of their work. *Journal of the Washington Academy of Sciences*, 101(3), 39–48.

Big Geo-data

Song Gao
Department of Geography, University of
California, Santa Barbara, CA, USA

Synonyms

Big georeferenced data; Big geospatial data;
Geospatial big data; Spatial big data

Definition/Introduction

Big geo-data is an extension to the concept of big data with emphasis on the geospatial component and under the context of geography or

geosciences. It is used to describe the phenomenon that large volumes of georeferenced data (including structured, semi-structured, and unstructured data) about various aspects of the Earth environment and society are captured by millions of environmental and human sensors in a variety of formats such as remote sensing imageries, crowdsourced maps, geotagged videos and photos, transportation smart card transactions, mobile phone data, location-based social media content, and GPS trajectories. Big geo-data is “big” not only because it involves a huge volume of georeferenced data but also because of the high velocity of generation streams, high dimensionality, high variety of data forms, the veracity (uncertainty) of data, and the complex interlinkages with (small) datasets that cover multiple perspectives, topics, and spatiotemporal scales. It poses grand research challenges during the life cycle of large-scale georeferenced data collection, access, storage, management, analysis, modeling, and visualization.

Theoretical Aspects

Geography has a long-standing tradition on the duality of research methodologies: the *law-seeking* approach and the *descriptive or explanatory* approach. With the increasing popularity of data-driven approaches in geography, a variety of statistical methods and machine learning methods have been applied in geospatial knowledge discovery and modeling for predictions. Miller and Goodchild (2015) discussed the major challenges (i.e., populations not samples, messy not clean data, and correlations not causality) and the role of theory in the data-driven geographic knowledge discovery and spatial modeling, with addressing the tensions between idiographic versus nomothetic knowledge in geography. Big geo-data is leading to new approaches to research methodologies in capturing complex spatiotemporal dynamics of the Earth and the society directly at multiple spatial and temporal scales instead of just snapshots. The data streams play a driving-force role in data-driven methods rather than a test or

calibration role behind the theory or models in conventional geographic analyses.

While data-driven science and predictive analytics evolve in geographic and provide new insights, sometimes it is still very challenging for humans to interpret the meanings of machine learning or analytical results or relate findings to underlying theory. To solve this problem, Janowicz et al. (2015) proposed a semantic cube to illustrate the need for semantic technologies and domain ontologies to address the role of diversity, synthesis, and definiteness in big data researches.

Social and Human Aspects

The emergence of big geo-data brings new opportunities for researchers to understand our socio-economic and human environments. In the journal of *Dialogues in Human Geography* (volume 3, Issue 3, November 2013), several human geographers and GIScience researchers discussed a series of theoretical and practical challenges and risks to geographic scholarship and raised a number of epistemological, methodological, and ethical questions related to the studies of big data in geography. With the advancements in location-awareness technology, information and communication technology, and mobile sensing technology, researchers employed emerging big geo-data for investigating the geographical perspective of human dynamics research within such contexts in the special issue on *Human Dynamics in the Mobile and Big Data Era* on the *International Journal of Geographical Information Science* (Shaw et al. 2016). By synthesizing multi-sources of big data, those researches can uncover interesting human behavioral patterns that are difficult or impossible to uncover with the traditional datasets. However, challenges still exist in the scarcity of demographics and cross-validation or getting the identity of individual behaviors rather than aggregated patterns. Moreover, the location-privacy concerns and discussions arise in both academic world and the society. There exist social tensions among big data accessibility and privacy protection.

Technical Aspects

Cloud computing technologies and their distributed deployment models offer scalable computing paradigms to enable big geo-data processing for scientific researches and applications. In the geospatial research world, cloud computing has attracted increasing attention as a way of solving data-intensive, computing-intensive, and access-intensive geospatial problems and challenges, such as supporting climate analytics, land-use and land-cover change analysis, and dust storm forecasting (Yang et al. 2017). Geocomputation facilitates fundamental geographical science studies by synthesizing high-performance computing capabilities with spatial analysis operations, with providing a promising solution to aforementioned geospatial research challenges.

There are a variety of big data analytics platforms and parallelized database systems emerging in the new era. They can be classified into two categories: (1) the massively parallel processing data warehousing systems like *Teradata* which are designed for holding large-scale structured data and support standard SQL queries and (2) the distributed file storage systems and cluster-computing framework like *Apache Hadoop* and *Apache Spark*. The advantages of Hadoop-based systems mainly lie in their high flexibility, scalability, low cost, and reliability for managing and efficiently processing a large volume of structured and unstructured datasets, as well as providing job schedules for balancing data, resources, and task loads. A MapReduce computation paradigm on Hadoop takes the advantages of divide-and-conquer strategy and improves the processing efficiency. However, big geo-data has its complexity on the spatial and temporal components and requires new analytical framework and functionalities compared with nonspatial big data. Gao et al. (2017) built a scalable Hadoop-based geoprocessing platform (GPHadoop) and ran big geo-data analytical functions to solve crowdsourced gazetteers harvesting problems. Recently, more efforts have been made in connecting traditional GIS analysis research community to the cloud computing research community for the next frontier of big geo-data analytics.

In one special issue on *big data* at the journal of *Annals of GIS* (volume 20, Issue 4, 2014), researchers further discussed several key technologies (e.g., cloud computing, high-performance geocomputation cyberinfrastructures) for dealing with quantitative and qualitative dynamics of big geo-data. Advanced spatiotemporal big data mining and geoprocessing methods should be developed by optimizing the elastic storage, balanced scheduling, and parallel computing resources in high-performance geocomputation cyberinfrastructures.

Conclusion

With the advancements in location-awareness technology and mobile distributed sensor networks, large-scale high-resolution spatiotemporal datasets about the Earth and the society become available for geographic research. The research on big geo-data involves interdisciplinary collaborative efforts. There are at least three research areas that require further work: (1) the systematic integration of various big geo-data sources in geospatial knowledge discovery and spatial modeling, (2) the development of advanced spatial analysis functions and models, and (3) the advancement of quality assurance issues on big geo-data. Finally, there will still be ongoing comparisons between data-driven and theory-driven research methodologies in geography.

Further Reading

- Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*, 61, 172–186.
- Janowicz, K., van Harmelen, F., Hendler, J., & Hitzler, P. (2015). Why the data train needs semantic rails. *AI Magazine*, Association for the Advancement of Artificial Intelligence (AAAI), pp. 5–14.
- Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *Geo Journal*, 80(4), 449–461.
- Shaw, S. L., Tsou, M. H., & Ye, X. (2016). Editorial: Human dynamics in the mobile and big data era. *International Journal of Geographical Information Science*, 30(9), 1687–1693.
- Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth*, 10(1), 13–53.

Big Georeferenced Data

► [Big Geo-data](#)

Big Geospatial Data

► [Big Geo-data](#)

Big Humanities Project

Ramon Reichert

Department for Theatre, Film and Media Studies,
Vienna University, Vienna, Austria

“Big Humanities” are a heterogenic field of research between IT, cultural studies, and humanities in general. Recently, because of higher availability of digital data, they gained even more importance. The term “Big Humanities Data” has prevailed due to the wider usage of the Internet, and it replaced the terms like “computational science” and “humanities computing,” which have been used since the beginning of the computer era in the 1960s. These terms were related mostly to the methodological and practical development of digital tools, infrastructures, and archives.

In addition to the theoretical explorations on science according to Davidson (2008), Svensson (2010), Anne et al. (2010) and Gold (2012), “Big Humanities Data” are divided into three trendsetting theoretical approaches, simultaneously covering the historical development and changes in the field of research according to the epistemological policy:

1. The usage of computers and digitalization of “primary data” within humanities and cultural studies are in the center of Digital humanities. On the one hand the digitization projects relate to the digitalized portfolios. On the other hand they relate to the computerized philology tools

for the application of secondary data or results. Even today these elementary methods of digital humanities are based on philological tradition, which sees the evidence-driven collection and management of data as the foundation of hermeneutics and interpretation. Beyond the narrow discussions about the methods, computer-based measuring within humanities and cultural studies claims the media-like postulates of objectivity within modern sciences. Contrary to the curriculum of text studies in the 50s and 60s within the “Humanities Computing” (McCarthy 2005) the research area of related disciplines has been differentiated and broadened to history of art, culture and sociology, media studies, technology, archaeology, history and musicology (Gold 2012).

2. According to the second phase, in addition to the quantitative digitalization of texts, the research practices are being developed in accordance with the methods and processes of production, analysis and modeling of digital research environments for work within humanities with digital data. This approach stands behind the enhanced humanities and tries to find new methodological approaches of qualitative application of generated, processed and archived data for reconceptualization of traditional research subjects. (Ramsey and Rockwell 2012, pp. 75–84).
3. The development from humanities 1.0 to humanities 2.0 (Davidson 2008, pp. 707–717) marks the transition from digital development of methods within “Enhanced Humanities” to the “Social Humanities” which use the possibility of web 2.0 to construct the research infrastructure. Social humanities use interdisciplinarity of scientific knowledge by making use of software for open access, social reading and open knowledge and by enabling online cooperative and collaborative work on research and development. On the basis of the new digital infrastructure of social web (hypertext systems, Wiki tools, Crowd funding software etc.) these products transfer the computer-based processes from the early phase of digital humanities into the network culture of the social sciences. Today it is

Blogging Humanities (work on digital publications and mediation in peer-to-peer networks) and *Multimodal humanities* (presentation and representation of knowledge within multimedia software environments) that stand for the technical modernization of academic knowledge (McPherson 2008). Because of them *Big Social Humanities* claims the right to represent paradigmatically alternative form of knowledge production. In this context one should reflect on the technical fundamentals of the computer-based process of gaining insights within the research of humanities and cultural studies while critically considering data, knowledge genealogy and media history in order to evaluate properly the understanding of a role in the context of digital knowledge production and distribution (Thaller 2012, pp. 7–23).

History of Big Humanities

Big Humanities have been considered only occasionally from the perspective of science and media history in the course of the last few years (Hockey 2004). Historical approach to the interdependent relation between humanities and cultural studies and the usage of computer-based processes relativize the aspiration of digital methods on the evidence and truth and support the argumentation that digital humanities were developed from a network of historical cultures of knowledge and media technologies with their roots in the end of the nineteenth century.

The relevant research literature of the historical context and genesis of Big Humanities is regarded as one of the first projects of genuine humanistic usages of computer a Concordance of Thomas of Aquino based on punch cards by Roberto Busa (Vanhoutte 2013, p. 126). Roberto Busa (1913–2011), an Italian Jesuit priest, is considered as a pioneer of Digital Humanities. This project enabled the achievement of uniformity in historiography of computational science in its early stage (Schischkoff 1952). Busa, who in 1949 developed the linguistic corpus of “Index Thomisticus” together with Thomas J. Watson,

the founder of IBM, (Busa 1951, 1980, pp. 81–90), is regarded a founder of the point of intersection between humanities and IT. The first digital edition on punch cards initiated a series of the following philological projects: “In the 60s the first electronic version of ‘Modern Language Association International Bibliography’ (MLAIB) came up, a specific periodical bibliography of all modern philologies, which could be searched through with a telephone coupler. The retrospective digitalization of cultural heritage started after that, having had ever more works and lexicons such as German vocabulary by Grimm brothers, historical vocabularies as the Krünitz or regional vocabularies” (Lauer 2013, p. 104).

At first, a large number of other disciplines and non-philological areas were formed such as literature, library, and archive studies. They had longer epistemological history in the field of philological case studies and practical information studies. Since the introduction of punch card methods, they have been dealing with quantitative and IT procedures for facilities of knowledge management. As one can see, neither the research question nor Busa’s methodological procedure have been without its predecessors, so they can be seen as a part of a larger and longer history of knowledge and media archeology. Sketch models of mechanical knowledge apparatus capable of combining information were found in the manuscripts of Suisse Archivar Karl Wilhelm Bührer (1861–1917, Bührer 1890, pp. 190–192). This figure of thought of flexible and modularized information unit was made to a conceptional core of mechanical data processing. The archive and library studies took part directly in the historical change of paradigm of information processing. It was John Shaw Billings, the doctor and later director of the National Medical Library, who worked further on the development of apparatus for machine-driven processing of statistical data, a machine developed by Hermann Hollerith in 1886 (Krajewski 2007, p. 43). Technology of punch cards traces its roots in technical pragmatics of library knowledge organization; even if only later – within the rationalization movement in the 1920s – the librarian working procedure

was automatized in specific areas. Other projects of data processing show that the automatized production of an *index* or a *concordance* marks the beginning of computer-based humanities and cultural studies for the lexicography and catalogue apparatus of libraries. Until the late 1950s, it was the automatized method of processing large text data with the punch card system after Hollerith procedure that stood in the center of the first applications/usages. The technical procedure of punch cards changed the lecture practice of text analysis by transforming a book into a database and by turning the linear-syntagmatic structure of text into a factual and term-based system. As early as 1951, the academic debate among the contemporaries started in academic journals. This debate saw the possible applications of the punch card system as largely positive and placed them into the context of economically motivated rationality. Between December 13 and 16, 1951, the *German Society for Documentation* and the *Advisory Board of German Economical Chamber* organized a working conference on the study of mechanization and automation of documentation process, which was enthusiastically discussed by philosopher Georgi Schischkoff. He talked about a “significant simplification and acceleration [...] by mechanical remembrance” (Schischkoff 1952, p. 290). The representatives of computer-based humanities saw in the “literary computing,” starting in the early 1950s, the first autonomous research area, which could provide an “objective analysis of exact knowledge” (Pietsch 1951). In the 1960s, the first studies in the field of computer linguistics concerning the automatized indexing of large text corpora appeared, publishing the computer-based analysis about word indexing, word frequency, and word groups.

The automatized evaluation procedure of texts for the editorial work within literary studies was described already in the early stages of “humanities computing” (mostly within its areas of “computer philology” and “computer linguistics”) on the ground of two discourse figures relevant even today. The first figure of discourse describes the achievements of the new tool usage with instrumental availability of data (“helping tools”); the other figure of discourse focuses on the

economical disclosure of data and emphasizes the efficiency and effectivity of machine methods of documenting. The media figure of automation was finally combined with the expectance that interpretative and subjective influences from the processing and analysis of information can be systematically removed. In the 1970s and 1980s, the computer linguistics was established as an institutionally positioned area of research with its university facilities, its specialist journals (*Journal of Literary and Linguistic Computing*, *Computing in the Humanities*), discussion panels (HUMANIST), and conference activities. The computer-based work in the historical-sociological research has its first large rise, but it remains regarded in the work reports less than an autonomous method, and it is seen mostly as a tool for critical text examination and as a simplification measure by quantifying the prospective subjects (Jarausch 1976, p. 13).

A sustainable media turn both in the field of production and in the field of reception aesthetics appeared with the application of standardized markup texts such as the *Standard Generalized Markup Language* established in 1986 and software-driven programs for text processing. They made available the additional series of digital modules, analytical tools, and text functions and transformed the text into a model of a database. The texts could be loaded as structured information and were available as (relational) databases. In the 1980s and 1990s, the technical development and the text reception were dominated by the paradigm of a database.

With the domination of the World Wide Web, the research and teaching practices changed drastically: the specialized communication experienced a lively dynamics through the digital network culture of publicly accessible online resources, e-mail distribution, chats, and forums, and it became largely responsive through the media-driven feedback mentality of rankings and voting. With its aspiration to go beyond the hierarchical structures of academic system through the reengineering of scientific knowledge, the Digital Humanities 2.0 made the ideals

of equality, freedom, and omniscience attainable again.

As opposed to its beginnings in the 1950s, the Digital Humanities today have also an aspiration to reorganize the knowledge of the society. Therefore, they regard themselves “both as a scientific as well as a socioutopistic project” (Hagner and Hirschi 2013, p. 7). With the usage of social media in the humanities and cultural studies, the technological possibilities and the scientific practices of *Digital Humanities* not only developed but they also brought to life new phantasmagoria of scientific distribution, quality evaluation, and transparency in the World Wide Web (Haber 2013, pp. 175–190). In this context, Bernhard Rieder and Theo Röhle identified five central problematic perspectives for the current “Digital Humanities” in their text from 2012 “five challenges.” These are the following: the temptation of objectivity, the power of visual evidence, black-boxing (fuzziness, problems of random sampling, etc.), institutional turbulences (rivaling service facilities and teaching subjects), and the claim of universality. Computer-based research is usually dominated by the evaluation of data so that some researchers see the advanced analysis within the research process even as a substitution for a substantial theory construction. That means that the research interests are almost completely data driven. This evidence-based concentration on the data possibilities can deceive the researcher to neglect the heuristic aspects of his own subject.

Since the social net is not only a neutral reading channel of research, writing, and publication resources without any power but also a governmental structure of power of scientific knowledge, the epistemological probing of social, political, and economic contexts of Digital Humanities includes also a data critical and historical questioning of its computer-based reformation agenda (Schreibmann 2012, pp. 46–58).

What did the usage of computer technology change for cultural studies and humanities on the basis of theoretical essentials? Computers did reorganize and accelerated the quantification and calculation process of scientific knowledge; they did entrench the metrical paradigm in the cultural

studies and humanities and promoted the hermeneutical-interpretative approaches with a mathematical formalization of the respective subject field. In addition to these epistemological shifts, the research practices within the Big Humanities have been shifted, since the research and development are seen as project related, collaborative, and network formed, and on the network horizon, they become the subject of research of network analysis. The network analysis itself has its goal to reveal the correlations and relation-patterns of digital communication of scientific networks and to declare the Big Humanities itself to the subject of reflections within a social constructivist actor-network-theory.

Further Reading

- Anne, B., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2010). *Digital humanities*. Cambridge, MA: MIT Press, 201(2). Online: http://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf.
- Bührer, K. W. (1890). Ueber Zettelnotizbücher und Zettelkatalog. *Fernschau*, 4, 190–192.
- Busa, R. (1951). *S. Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*. Milano: Bocca.
- Busa, R. (1980). The annals of humanities computing: The index Thomisticus. *Computers and the Humanities*, 14(2), 83–90.
- Davidson, C. N. (2008). Humanities 2.0: Promise, perils, predictions. *Publications of the Modern Language Association (PMLA)*, 123(3), 707–717.
- Gold, M. K. (Ed.). (2012). *Debates in the digital humanities*. Minneapolis: University of Minnesota Press.
- Haber, P. (2013). ‘Google Syndrom’. Phantasmagorien des historischen Allwissens im World Wide Web. *Zürcher Jahrbuch für Wissensgeschichte*, 9, 175–190.
- Hagner, M., & Hirschi, C. (2013). Editorial Digital Humanities. *Zürcher Jahrbuch für Wissensgeschichte*, 9, 7–11.
- Hockey, S. (2004). History of humanities computing. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities*. Oxford: Blackwell.
- Jarausch, K. H. (1976). Möglichkeiten und Probleme der Quantifizierung in der Geschichtswissenschaft. In: ders., *Quantifizierung in der Geschichtswissenschaft. Probleme und Möglichkeiten* (pp. 11–30). Düsseldorf: Droste.
- Krajewski, M. (2007). In Formation. Aufstieg und Fall der Tabelle als Paradigma der Datenverarbeitung. In D. Gugerli, M. Hagner, M. Hampe, B. Orland, P. Sarasin, & J. Tanner (Eds.), *Nach Feierabend. Zürcher Jahrbuch für Wissenschaftsgeschichte* (Vol. 3, pp. 37–55). Zürich/Berlin: Diaphanes.
- Lauer, G. (2013). Die digitale Vermessung der Kultur. Geisteswissenschaften als Digital Humanities. In H. Geiselberger & T. Moorstedt (Eds.), *Big Data. Das neue Versprechen der Allwissenheit* (pp. 99–116). Frankfurt/M: Suhrkamp.
- McCarty, W. (2005). *Humanities computing*. London: Palgrave.
- McPherson, T. (2008). Dynamic vernaculars: Emergent digital forms in contemporary scholarship. Lecture presented to HUMLab Seminar, Umeå University, 4 Mar. <http://stream.humlab.umu.se/index.php?streamName=dynamiCVernaculars>.
- Pietsch, E. (1951). Neue Methoden zur Erfassung des exakten Wissens in Naturwissenschaft und Technik. *Nachrichten für Dokumentation*, 2(2), 38–44.
- Ramsey, S., & Rockwell, G. (2012). Developing things: Notes toward an epistemology of building in the digital humanities. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 75–84). Minneapolis: University of Minnesota Press.
- Rieder, B., & Röhle, T. (2012). Digital methods: Five challenges. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 67–84). London: Palgrave.
- Schischkoff, G. (1952). Über die Möglichkeit der Dokumentation auf dem Gebiete der Philosophie. *Zeitschrift für Philosophische Forschung*, 6(2), 282–292.
- Schreibman, S. (2012). Digital humanities: Centres and peripheries. In: M. Thaller (Ed.), *Controversies around the digital humanities (Historical social research, Vol. 37(3), pp. 46–58)*. Köln: Zentrum für Historische Sozialforschung.
- Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities Quarterly (DHQ)*, 4(1). Online: <http://www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html>.
- Thaller, M. (Ed.). (2012). Controversies around the digital humanities: An agenda. *Computing Historical Social Research*, 37(3), 7–23.
- Vanhoutte, E. (2013). The gates of hell: History and definition of digital | humanities. In M. Terras, J. Tyham, & E. Vanhoutte (Eds.), *Defining digital humanities* (pp. 120–156). Farnham: Ashgate.

Big O Notation

► Algorithmic Complexity

Big Variety Data

Christopher Nyamful¹ and Rajeev Agrawal²

¹Department of Computer Systems Technology,
North Carolina A&T State University,
Greensboro, NC, USA

²Information Technology Laboratory, US Army
Engineer Research and Development Center,
Vicksburg, MS, USA

Introduction

Massive data generated from daily activities which has accumulated over the years may contain valuable information and insights which can be leveraged to assist in decision-making for greater competitive advantage. Consider data from weather, traffic control, satellite imagery, geography, and social media data to daily sales figures that have inherent patterns that, if discovered, can be used to forecast likely future occurrences. The huge amount of user-generated data is unstructured; its unstructured content has no conceptual data-type definition. They are typically stored as files, such as word documents, PowerPoint presentations, photos, videos, web pages, blogs such as tweets, and Facebook posts.

Unstructured data is the most common, and people use it every day. For example, the use of video surveillance has increased, likewise satellite-based remote sensing and aerial photography of both optical and multispectral imagery. The smartphone is also a good example of how a mobile device produces an additional variety of data sources that is captured for reuse. Most of the files that organizations want to keep are usually image based. Industries and government regulations require a significant portion, if not all unstructured data be stored for long-term retention and access. This data must be appropriately classified and managed for future analysis, search, and discovery. The notion of using data variety entails the idea of using multiple sources of data to help understand and solve a problem.

Large data sets in a big data environment are made up of varying data types. Data can be classified as structured, semi-structured, and unstructured based on how it is stored and analyze. Semi-structured data is a kind of structured data that is not raw or strictly typed. They don't have any underlying data model, hence cannot be associated with any relational database. The web provides numerous examples of semi-structured data such as hypertext markup language (html) and extensible markup language (xml). Structured data is organized in a strict format of rows and columns. It makes use of data model which determines the schema for the structured data. Data types under structured data can be organized by index and queried in various ways to yield required results. Relational database management systems are used to analyze and manage structured data.

About 90 percent of big data is highly unstructured (Cheng et al. 2012). The primary concern of businesses and organizations is how to manage unstructured data, because they form the bulk part of data received and processed. Unstructured data requires a significant amount of storage space. Storing large collection of digital files and streaming videos have become common in today's era of big data. For instance, Youtube receives one billion unique users every day, and 100 h of video is uploaded each minute ("YouTube Data Statistics" 2015). Clearly, there is a massive increase in video file storage requirements, in terms of capacity and IOPS.

As data sets increase in both structured and unstructured forms, analysis and management get more diverse. The real-time component of big data environment poses a great challenge. The use of web commercials based on users' purchase and search history requires real-time analytics. In order to effectively manage these huge quantities of unstructured data, a high IOPS performance storage environment is required. A wide range of technologies and techniques have been developed to analyze, manipulate, aggregate, and visualize big data.

Current Systems

A scale-out network-attached storage (NAS) is a network storage system used to simplify storage management through a centralized point of control. It pools multiple storage nodes in a cluster. This cluster performs NAS processes as a single entity. Unlike traditional NAS, scale-out NAS provides the capability of nodes or heads to be added as processing power demands. It supports file I/O-intensive applications and scales to petabytes of data. Scale-out NAS provides the platform for a flexible management of diverse data types in big data. It's characterized with moderate performance and availability to produce a complete system with better aggregate computing power and availability. The use of gigabit Ethernet allows scale-out NAS to be deployed over a wide geographical area and still maintain high throughput.

Most storage vendors are showing more interest in scale-out NAS to deal with the challenges of big data with media-rich files – unstructured data. Storage vendors differ in a way they architect scale-out network-attached storage. EMC Corporation offers Isilon OneFS scale-out NAS to its clients. Isilon provides the capabilities to meet big data challenges. It comes with a specialized operating system known as OneFS. Isilon OneFS consolidate file system, volume manager, and Redundant Array of Independent Disk (RAID) into a unified software layer and a single file system that is distributed across all nodes in the cluster. EMC scale-out NAS simplifies storage infrastructure and reduces cost by consolidating unstructured data sets and large-scale files, eliminating storage silos. It provides massive scalability for unstructured big data storage needs, ranging from 16 TB to 20 TB capacity per cluster. Isilon's native Hadoop Distributed File System can be leveraged to support Hadoop analytics on both structured and unstructured data. EMC Isilon moderate performance can reach up to 2.6 million file operations per second with over 200 gigabyte per second of aggregate throughput to support the demands posed by big data workloads. Other

storage vendors such as IBM, NetApp Inc., Hitachi Data systems, Hewlett Packard (HP), Dell Inc., and among others offer scale-out NAS to address the unstructured big data needs.

Object-based storage system(OSD) offers an innovative platform for storing and managing unstructured data. It stores data in the form of objects based on its content and other attribute. An object has a variable length and can be used to store any type of data. It provides an integrated solution that supports file-, block-, and object-level access to storage devices. Object-based storage devices organize and stores unstructured data such as movies, photos, and documents as objects. OSD uses flat address space to store data and uses a unique identifier to access that data. The use of the unique identifier eliminates the need to know specific location of a data object. Each object is associated with an object ID, generated by a special hash function and guarantees each object is uniquely identified. The object is also composed of data, attributes, and rich metadata. The metadata keeps track of the object content and makes access, discovery, distribution, and retention much more feasible. Object storage brings structure to unstructured data, making it easier to store, protect, secure, manage, organize, search, sync, and share file data. The great features provided by OSD allow organizations to leverage a single storage investment for a variety of workloads.

Hitachi Data Systems offers object-based storage solution that treats data files, metadata, and file attributes as a single object that is tracked and retained among a variety of storage tiers. They provide multiple fields for metadata so that different users and application can use their own metadata and tag without conflict. EMC Atmos storage system is designed to support object-based storage for unstructured data such as videos and pictures. Atmos integrates massive scalability with high performance to address challenges associated with vast amount of unstructured data. It enhances operational efficiency by distributing content automatically based on business policy. Atmos also provides data services such as replication, deduplication, and compression. Atmos

multitenancy feature allows multiple applications to be processed from the same infrastructure.

Distributed Systems

Apache Hadoop project has developed open-source software for reliable, scalable, and efficient distributed computing. Hadoop Distributed File System (HDFS) is a distributed file system that stores data on low-cost machines, providing high aggregate bandwidth across the cluster (Shvachko et al. 2010). HDFS stores huge files across multiple nodes. It ensures reliability by replicating data across multiple hosts in a cluster. HDFS is composed of two agents, namely, NameNode and DataNode. NameNode is responsible for managing metadata and DataNode manages data input/output. Each DataNode serves up blocks of data over the network using a block protocol specific to HDFS and uses the standard IP network for communication. HDFS has master/slave architecture. Input files distributed across the cluster are automatically split into even-sized chunks which are managed by different nodes in the cluster. It ensures scalability and availability. For example, Yahoo uses Hadoop to manage 25 PB of enterprise data in 25,000 servers.

Hadoop uses distributed architecture known as MapReduce for mapping tasks to servers for processing. Amazon Elastic MapReduce (Amazon EMR) uses Hadoop to analyze and processes huge amount of data, both structured and unstructured data. It achieves this by distributing workloads across virtual servers running in the Amazon cloud. Amazon EMR simplifies the use of Hadoop and big data-intensive applications. Amazon Elastic Compute Cloud (EC2) is being used by various organizations to process vast amount of unstructured data. The New York Times rents 100 virtual machines to convert 11 million scanned articles to PDFs.

The relational database is not able to support the vast variety of unstructured data which are being received from all sources of digital activities. NoSQL databases are now being deployed to address some of the big unstructured data challenges. NoSQL represents a class of data management technologies designed to address high-

volume, high variety, and high velocity data. Comparatively, they are more scalable and support superior performance. MongoDB is a cross platform NoSQL database that is designed to overcome the limitations of the traditional database. MongoDB is optimized for efficiency, and its features include:

1. Scale-out architecture, instead of expensive monolithic architecture
2. Supports large volumes of structured, semi-structured, and structured data
3. Agile sprints, quick iteration, and frequent code pushes
4. Flexibility in use of object-oriented programming

An alternative area for addressing big data-related issues is the cloud. The emergence of cloud computing has eased IT burdens of most organizations. Storage and analysis can be outsourced to the cloud. In the era of big data, the cloud offers a potential self-service consumption model for data analytics. Cloud computing allows organizations and individuals to obtain IT resources as a service. Cloud computing offers it services according to several fundamental model – Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The Amazon Compute Cloud (EC2) is an example of IaaS that provides on-demand services in the cloud for clients. IaaS allow access to its infrastructure for client application to be deployed onto. It shares and manages a pool of configurable and scalable resources such as network and storage servers. Google App Engine is an example of PaaS that allows clients to develop and run their software solutions on the cloud platform. PaaS guarantees high computing platforms to meet the different workload from clients. Cloud providers manage required computing infrastructure and software support service to clients. SaaS model allows clients to use provided application to meet their business needs. The cloud uses its multitenant feature to accommodate a large number of users. Flickr, Amazon, and Google docs are great examples of SaaS. Both cloud computing and big data analytics are extension of virtualization technologies. Virtualization

abstract physical resources such as storage, compute, and network and make them appear as logical resources. Cloud infrastructure is usually built on virtualized data center by providing resource pooling. Organizations are deploying virtualization technique across data centers to optimize their use.

Conclusion

Big variety data is on the rise and touches all areas of life, especially with the high-degree usage of the Internet. Methods for simplifying big variety data in terms of storage, integration, analysis, and visualization are complex. Current storage systems, to a considerable extent, are addressing some big data-related issues. A high-performance system to ensure maximum data transfer rate and analysis has become a research focus. Current systems can be improved in the near future to handle efficiently the vast amount of unstructured data in the big data environment.

Further Reading

- Cheng, Y., Qin, C., & Rusu, F. (2012). *GLADE: Big data analytics made easy*. Paper presented at the Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, Scottsdale.
- Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). *The hadoop distributed file system*. Paper presented at the 2010 I.E. 26th Symposium on Mass Storage Systems and Technologies (MSST), IEEE.
- YouTube Data Statistics. (2015). Retrieved 15 Jan 2015, from <http://www.youtube.com/yt/press/statistics.html>.

its basis in the study of genotypes and phenotypes, the bioinformatics domain is extensive, encompassing genomics, metagenomics, proteomics, pharmacogenomics, and metabolomics. With advances in IT-supported data storage and management, very large data sets (Big Data) have become available from diverse sources at greatly accelerated rates, providing unprecedented opportunities to engage in increasingly more sophisticated biological data analytics.

Although the formal study of biology has its origins in the seventeenth century CE, the application of computer science to biological research is relatively recent. In 1953, the Watson and Crick published the DNA structure. In 1975, Sanger and the team of Maxam and Gilbert independently developed DNA sequences. In 1980, the US Supreme Court ruled that patents on genetically modified bacteria are allowed. This ruling made pharmaceutical applications a primary motive for human genomic research: the profits from drugs based on genomic research could be enormous. Genomic research became a race between academe and the commercial sector. Academic consortia, comprising universities and laboratories, rushed to place their gene sequences in the public domain to prevent commercial companies from applying for patents on those sequences. The 1980s also saw the sequencing of human mitochondrial DNA (1981; 16,589 base pairs) and the Epstein-Barr virus genome (1984; 172,281 base pairs). In 1990, the International Human Genome project was launched with a projected 15-year duration. On 26 June 2006, the draft of the human genome was published, reflecting the successful application of informatics to genomic research.

Bioinformatics

Erik W. Kuiler
George Mason University, Arlington, VA, USA

Background

Bioinformatics constitute a specialty in the informatics domain that applies information technologies (IT) to the study of human biology. Having

Biology, Information Technology, and Big Data Sets

For much of its history, biology was considered to be a science based on induction. The introduction of computer science to the study of biology and the evolution of data management from mainframe- to cloud-based computing provides the impetus for the progression from strictly observation-based biology to bioinformatics. Current

relational database management systems, originally developed to support business transaction processes, were not designed to support data sets of one or more petabytes (10^{15} bytes) or exabytes (10^{18} bytes). Big biological data sets are likely to contain structured as well as unstructured data, including structured quantitative data, images, and unstructured text data. Advances in Big Data storage management and distribution support large genomic data mappings; for example, nucleotide databases may contain more than 6×10^{11} base pairs, and a single sequenced human genome may be approximately 140 gigabytes in size. With the adoption and promulgation of Big Data bioinformatics, human genomic data can be embedded in electronic health records (EHR), facilitating, for example, individualized patient care.

Bioinformatics Transparency and International Information Exchange

The number of genomic projects is growing. Started in 2007 as a pilot project, the Encyclopaedia of DNA Elements (ENCODE) is an attempt to understand the functions of the human genome. The results produced by projects such as ENCODE are expected generate more genomics-focused research projects. The National Human Genome Research Institute (NHGRI), a unit of the National Institutes of Health (NIH), provides a list of educational resources on its website. The volume of bioinformatics-generated research has led to the development of large, online research databases, of which PubMed, maintained by the US National Library of Medicine, is just one example.

Genomic and biological research is an international enterprise, and there is a high level of transnational collaboration to assure data sharing. For example, the database of nucleic acid sequences is maintained by a consortium comprising institutions from the US, UK, and Japan. The UK-based European Bioinformatics Institute (EBI) maintains the European Nucleotide Archive (ENA). The US National Center for Biotechnology Information maintains the International

Nucleotide Sequence Database Collaboration. The Japanese National Institute of Genetics supports the DNA Data Bank of Japan and the Center for Information Biology. To ensure synchronized coverage, these organizations share information on a regular basis. For organizations that manage bioinformatics databases, providing browsers to access and explore their contents has become a de facto standard; for example, the US National Center for Biotechnology Information offers ENTREZ to execute parallel searches in multiple databases.

Translational Bioinformatics

Bioinformatics conceptualize biology at the molecular level and organize data to store and manage them efficiently for research and presentation. Translational bioinformatics focus on the transformation of biomedical data into information to support, for example, the development of new diagnostic techniques, clinical interventions, or new commercial products and services. In the pharmacological sector, translational bioinformatics provide the genetic data necessary to repurpose existing drugs. Genetic data may also prove useful in differentiating drug efficacy based on gender or age. In effect, translational bioinformatics enable bi-directional data sharing between the research laboratory and the medical clinic. Translational bioinformatics enable hospitals to develop clinical diagnostic support capabilities that incorporate correlations between individual genetic variations and clinical risk factors, disease presentations, or responses to treatment.

Bioinformatics and Health Informatics

Translational bioinformatics-generated data can be incorporated in EHRs that are structured to contain both structured and unstructured data. For example, with an EHR that complies with the Health Level 7 Consolidated Clinical Document Architecture (HL7 C-CDA), it is possible to share genetic data (personal genomic data), X-ray

images, and diagnostic data, as well as a clinician's free-form notes. Because EHR data are usually entered to comply with predetermined standards, there is substantially less likelihood of error in interpretation or legibility. Combined, Health Information Exchange (HIE) and EHR provide the foundation for biomedical data sharing. HIE operationalizes the Meaningful Use (MU) provisions of the Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted as Titles IV and XIII the American Recovery and Reinvestment Act (ARRA) of 2009, by enabling information sharing among clinicians, patients, payers, care givers, federal and state agencies.

Biomedical Data Analytics

Biomedical research depends on quantitative data as well as unstructured text data. With the availability of Big Data sets, selecting the appropriate analytical model depends on the kind of analysis we plan to undertake, the kinds of data we have, and the size of the data set. Data mining models, such as artificial neural networks, statistics-based models, and Bayesian models that focus on probabilities and likelihoods support predictive analytics. Classification models are useful in determining the category where an individual object belongs based on identifiable properties. Clustering models are useful for identifying population subsets based on shared parameters. Furthermore, advances in computer science have also led to the development and analysis of algorithms, not only in terms of complexity but also in terms of performance. Induction-based algorithms are useful in unsupervised learning settings; for example, text mining or topic analysis for the purpose of exploration, where we are not trying to prove or disprove a hypothesis but are simply exploring a body of documents for lexical clusters and patterns. In contrast, deduction-based algorithms can be useful in supervised learning settings, where there are research questions to be answered and hypotheses to be tested. In the health domain, random clinical trials (RCT) are archetypal examples of hypothesis-based model development.

Challenges and Future Trends

To remain epistemically viable, bioinformatics, like health informatics, require the capabilities to ingest, store, and manage Big Data sets. However, these capabilities are still in their infancy. Similarly, data analytics tools may not be sufficiently efficient to support Big Data exploration in a timely manner. Because personal genomic information can now be used in EHRs, translational bioinformatics, like health informatics, must incorporate stringent anonymization controls. Bioinformatics are beginning to develop computational models of disease processes. These can prove beneficial not only to the development or modification of clinical diagnostic protocols and interventions but also for epidemiology and public health. In academe, bioinformatics are increasingly accepted as multidiscipline programs, drawing their expertise from biology, computer science, statistics, and medicine (translational bioinformatics).

Based on the evolutionary history of IT development, dissemination, and acceptance, it is likely that IT-based technical issues in Big Data-focused bioinformatics will be addressed and that the requisite IT capabilities will become available over time. However, there are a number of ethical and moral issues that attend the increasing acceptance of translational bioinformatics-provided information in the health domain. For example, is it ethical or moral for a health insurance provider to deny coverage based on personal genomics? Also, is it appropriate, from a public policy perspective, to use bioinformatics-generated data to institute eugenic practices, even if only de facto, to support the social good? As a society it behooves us to address questions such as these.

Further Reading

- Butte, A. (2008). Translational bioinformatics: Coming of age. *Journal of the American Medical Informatics Association*, 15(6), 709–714.
- Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, 33(7), 1139–1147.

- Kumari, D., & Kumari, R. (2014). Impact of biological big data in bioinformatics. *International Journal of Computer Applications*, 10(11), 22–24.
- Maajo, V., & Kulikowski, C. A. (2003). Bioinformatics and medical informatics: Collaborations on the road to genomic medicine? *Journal of the American Medical Informatics Association*, 10(6), 515–522.
- Ohno-Machado, L. (2012). Big science, big data, and the big role for biomedical informatics. *Journal of the American Medical Informatics Association*, 19(e1), e1.
- Shah, N. H., & Tenebaum, J. D. (2012). The coming of age of data-driven medicine: Translational bioinformatics' next frontier. *Journal of the American Medical Informatics Association*, 19, e1–e2.

Biomedical Data

Qinghua Yang¹ and Fan Yang²

¹Department of Communication Studies, Texas Christian University, Fort Worth, TX, USA

²Department of Communication Studies, University of Alabama at Birmingham, Birmingham, AL, USA

Thanks to the development of modern data collection and analytic techniques, biomedical research generates increasingly large amounts of data in various formats and at all levels, which is referred to as big data. Big data is a collection of data sets, which are large in volume and complex in structure. To illustrate, the data managed by America's leading healthcare provider Kaiser is 4,000 times more than the amount of information stored in the Library of Congress. As to data structure, the range of nutritional data types and sources make it really difficult to normalize. Such volume and complexity of big data make it difficult to be processed by traditional data analytic techniques. Therefore, to further knowledge and uncover hidden value, there is an increasing need to better understand and mine biomedical big data by innovative techniques and new approaches, which requires interdisciplinary collaborations involving data providers and users (e.g., biomedical

researchers, clinicians, and patients), data scientists, funders, publishers, and librarians.

The collection and analysis of big data in biomedical area have demonstrated its ability to enable efficiencies and accountability in health care, which provides strong evidence for the benefits of big data usage. Electronic health records (EHRs), an example of biomedical big data, can provide timely data for assisting monitoring of infectious diseases, disease outbreaks, and chronic illnesses, which could be particularly valuable during public health emergencies. By collecting and extracting data from EHRs, public health organizations and authorities could receive extraordinary amount of information. By analyzing the massive data from EHRs, public health researchers could conduct comprehensive observational studies with uncountable patients who are treated in real clinical settings over years. Disease progress, clinical outcomes, treatment effectiveness, and public health intervention efficacies can also be studied by analyzing EHRs data, which may influence public health decision-making (Hoffman and Podgurski 2013).

As a crucial juncture of addressing the opportunities and challenges presented by biomedical big data, the National Institutes of Health (NIH) has initiated a Big Data to Knowledge (BD2K) initiative to maximize the use of biomedical big data. BD2K, a response to the Data and Informatics Working Groups (DIWG), focuses on enhancing:

- (a) the ability to locate, access, share, and apply biomedical big data,
- (b) the dissemination of data analysis methods and software,
- (c) the training in biomedical big data and data science,
- (d) the establishment of centers of excellence in data science (Margolis et al. 2014)

First, BD2K initiative fosters the emergence of data science as a discipline relevant to biomedicine by developing the solutions to specific high-

need challenges confronting the research community. For instance, the Centers of Excellence in Data Science initiated the first BD2K Funding Opportunity to test and validate new ideas in data science. Second, BD2K aims to enhance the training of methodologists and practitioners in data science by improving their skills in demand under the data science “umbrella,” such as computer science, mathematics, statistics, biomedical informatics, biology, and medicine. Third, given the complex questions posed by the generation of large amounts of data requiring interdisciplinary teams, BD2K initiative facilitates the development of investigators in all parts of the research enterprise for interdisciplinary collaboration to design studies and perform subsequent data analyses (Margolis et al. 2014).

Besides these promotive initiatives proposed by national research institutes, such as NIH, great endeavors in improving biomedical big data processing and analysis have also been made by biomedical researchers and for-profit organizations. National cyberinfrastructure has been suggested by biomedical researchers as one of the systems that could efficiently handle many of big data challenges facing the medical informatics community. In the United States, the national cyberinfrastructure (CI) refers to an existing system of research supercomputer centers and high-speed networks that connect them (LeDuc et al. 2014). CI has been widely used by physical and earth scientists, and more recently biologists, yet little used by biomedical researchers. It has been argued that more comprehensive adoption of CI could facilitate many challenges in biomedical area. One example of innovative biomedical big data technique provided by for-profit organizations is GENALICE MAP, a next-generation sequencing (NGS) DNA processing software launched by a Dutch Software Company GENALICE. Processing biomedical big data one hundred times faster than conventional data analytic tools, MAP demonstrated robustness and spectacular performance and raised the NGS data processing and analysis to a new level.

Challenges

Despite the opportunities brought by biomedical big data, certain noteworthy challenges also exist. First, to use big biomedical data effectively, it is imperative to identify the potential sources of healthcare information and to determine the value of linking them together (Weber et al. 2014). The “bigness” of biomedical data sets is multidimensional: some big data, such as EHRs, provide depth by including multiple types of data (e.g., images, notes, etc.) about individual patient encounters; others, such as claims data, provide longitudinality, which refers to patients’ medical information over a period of time. Moreover, social media, credit cards, census records, and a various number of other types of data can help assemble a holistic view of a patient and shed light on social and environmental factors that may be influencing health.

The second technical obstacle in linking big biomedical data results from the lack of a national unique patient identifier (UPI) in the United States (Weber et al. 2014). To address the absence of a UPI to enable precise linkage, hospitals and clinics have developed sophisticated probabilistic linkage algorithms based on other information, such as demographics. By requiring enough variables to match, hospitals and clinics are able to reduce the risk of linkage errors to an acceptable level even though two different patients share the same characteristics (e.g., name, age, gender, zip code). In addition, the same techniques used to match patients across different EHRs can be extended to data sources outside of health care, which is an advantage of probabilistic linkage.

Third, besides the technical challenges, privacy and security concerns turn to be a social challenge in linking biomedical big data (Weber et al. 2014). As more data are linked, they become increasingly more difficult to be de-identified. For instance, although clinical data from EHRs offer considerable opportunities for advancing clinical and biomedical research, unlike most other forms of biomedical research

data, clinical data are typically obtained outside of traditional research settings and must be converted for research use. This process raises important issues of consent and protection of patient privacy (Institute of Medicine 2009). Possible constructive responses could be to regulate legality and ethics, to ensure that benefits outweigh risks, to include patients in the decision-making process, and to give patients control over their data. Additionally, changes in policies and practices are needed to govern research access to clinical data sources and facilitate their use for evidence-based learning in healthcare. Improved approaches to patient consent and risk-based assessments of clinical data usage, enhanced quality and quantity of clinical data available for research, and new methodologies for analyzing clinical data are all needed for ethical and informed use of biomedical big data.

Cross-References

- [Biometrics](#)
- [Data Sharing](#)
- [Health Informatics](#)

Further Reading

- Hoffman, S., & Podgurski, A. (2013). Big bad data: Law, public health, and biomedical databases. *The Journal of Law, Medicine & Ethics*, 41(8), 56–60.
- Institute of Medicine. (2009). *Beyond the HIPAA privacy rule: Enhancing privacy, improving health through research*. Washington, DC: The National Academies Press.
- LeDuc, R., Vaughn, M., Fonner, J. M., Sullivan, M., Williams, J. G., Blood, P. D., et al. (2014). Leveraging the national cyberinfrastructure for biomedical research. *Journal of the American Medical Informatics Association*, 21(2), 195–199.
- Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., et al. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: Capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, 21(6), 957–958.
- Weber, G., Mandl, K. D., & Kohane, I. S. (2014). Finding the missing link for big biomedical data. *Journal of American Medical Association*, 331(4), 2479–2480.

Biometrics

Jörgen Skågeby

Department of Media Studies,
Stockholm University, Stockholm, Sweden

Biometrics refers to measurable and distinct (preferably unique) biological, physiological, or behavioral characteristics. Stored in both commercial and governmental biometric databases, these characteristics are subsequently used to identify and/or label individuals. This entry summarizes common forms of biometrics, their different applications and the societal debate surrounding biometrics, including its connection to big data, as well as its potential benefits and drawbacks.

Typical applications for biometrics include identification in its own right, but also as a measure to verify access privileges. As mentioned, biometrical technologies rely on physiological or, in some cases, behavioral characteristics. Some common physiological biometric identifiers are eye retina and iris scans, fingerprints, palm prints, face recognition and DNA. Behavioral biometric identifiers can include typing rhythm (keystroke dynamics), signature recognition, voice recognition, or gait. While common sites of use include border controls, education, crime prevention and health care, biometrics are also increasingly deployed in consumer devices and services, such as smartphones (e.g., fingerprint verification in the iPhone 5 and onwards, as well as facial recognition in Samsung's Galaxy phones) and various web services and applications making use of keystroke dynamics. Although some biometric technologies have been deployed for over a century, the recent technological development has spurred a growing interest in a wider variety of biometrical technologies and their respective potential benefits and drawbacks. As such, current research on biometrics is not limited to technical or economical details. Rather, political, ontological, social, and ethical aspects and implications are now often considered in cohort with technical advances. As a consequence, questions around the convergence of the biological

domain and the informational domain are given new relevance. Because bodily individualities are increasingly being turned into code/information, digital characteristics (e.g., being easily transferrable, combinable, searchable, and copyable) are progressively applicable to aspects of the corporeal, causing new dilemmas to emerge. Perhaps not surprisingly then, both the suggested benefits and drawbacks with biometric technologies and the connected big data repositories tie into larger discussions on privacy in the digital age.

One of the most commonly proposed advantages of biometrics is that it provides a (more or less) unique identity verifier. This, in turn, makes it harder to forge or steal identities. It is also argued that such heightened fidelity in biometrical measures improves the general level of societal security. Proponents will also argue that problems caused by lost passports, identity cards or driver's licenses, as well as forgotten passwords are virtually eliminated. More ambivalent advantages of biometrics include the possibility to automatically and unequivocally tie individuals to actions, geographical positions, and moments in time. While such logs can certainly be useful in some cases, they also provide opportunities for more pervasive surveillance.

Consequently, the more ambivalent consequences, combined with a more careful consideration of the risks connected to biometric data retention, have generated many concerns about the widened deployment of biometric technologies. For example, biometric databases and archives are often regarded as presenting too many risks. Even though the highest security must be maintained to preserve the integrity of these digital archives, they can still be hacked or used (by both governmental and commercial actors as well as individuals who have access to the information in their daily work) in ways not anticipated, sanctioned, or legitimized. There is also a question of continuously matching physical bodies to the information stored in databases. Injuries, medical conditions, signs of aging, and voluntary bodily modifications may cause individuals to end up without a matching database entry, effectively causing citizens and users to be locked out from their own identity. Disbelievers

will also argue that there is still a prevailing risk of stolen or forged identities. While the difficulty of counterfeiting biometric data can be seen as directly related to the sophistication of the technology used to authenticate biometric data, a general argument made is that of a "technological balance of terror." That is, as biometric technologies develop to become more sophisticated and sensitive, so will the technologies capable of forging identities. More so however, the risk of stolen proofs of identity still presents a very real risk, particularly when conducted in a networked digital environment. Digital files can be easily copied and unlike a physical lock, which can be exchanged if the key is stolen, once such a digital file (of, e.g., a fingerprint or iris scan) is stolen it may present serious and far-reaching repercussions. Thus, detractors argue that even with the proper policies in place the related incidents can be problematic. Furthermore, it will become even harder to foresee the potential problems and misuses of biometrics and as a consequence the public trust in biometric archives and technologies will be hard to maintain.

On a larger scale, the potential cooperation between commercial actors and governments has become a cause for concern for critics. They argue that practices previously regulated to states (and even then questionable) have now been adopted by commercial actors turning biometrical technologies into potential tools for general surveillance. Critics argue that instead of regulating the collection of biometric data to those who are convicted, registration of individuals has now spread to the general population. As such, opponents argue that under widespread biometric application all citizens are effectively treated as potential threats or suspected criminals.

In summary, biometrics refers to ways of using the human body as a verification of identity. Technologies making use of biometric identifiers are becoming increasingly common and will likely be visible in a growing number of applications in the everyday lives of citizens. Due to the ubiquitous collection and storage of biometric data through an increasingly sophisticated array of pervasive technologies and big data repositories, many critical questions are raised around the ontological,

social, ethical, and political consequences of their deployment. Larger discussions of surveillance, integrity, and privacy put biometric technologies and databases in a position where public trust will be a crucial factor in its overall success or failure.

Further Reading

- Ajana, B. (2013). *Governing through biometrics: The biopolitics of identity*. London: Palgrave Macmillan.
- Gates, K. (2011). *Our biometric future*. New York: New York University Press.
- Magnet, S. (2011). *When biometrics fail*. Durham: Duke University Press.
- Payton, T., & Claypoole, T. (2014). *Privacy in the age of big data*. Lanham: Rowman & Littlefield.

Biosurveillance

Ramón Reichert

Department for Theatre, Film and Media Studies,
Vienna University, Vienna, Austria

Internet biosurveillance, or *Digital Disease Detection*, represents a new paradigm of *Public Health Governance*. While traditional approaches to health prognosis operated with data collected in the clinical diagnosis, Internet biosurveillance studies use the methods and infrastructures of *Health Informatics*. That means, more precisely, that they use unstructured data from different web-based sources and targets using the collected and processed data and information about changes in health-related behavior. The two main tasks of the Internet biosurveillance are (1) the early detection of epidemic diseases, biochemical, radiological, and nuclear threats (Brownstein et.al. 2009) and (2) the implementation of strategies and measures of sustainable governance in the target areas of health promotion and health education (Walters et al. 2010). Biosurveillance has established itself as an independent discipline in the mid-1990s, as military and civilian agencies began to get interested in automatic monitoring systems. In this context, the *biosurveillance program* of the

Applied Physics Laboratory of Johns Hopkins University has played a decisive and pioneering role (Burkom et al. 2008).

The Internet biosurveillance uses the accessibility to data and analytic tools provided by digital infrastructures of social media, *participatory sources*, and *non-text-based sources*. The structural change generated by digital technologies, as main driver for Big Data, offers a multitude of applications for sensor technology and biometrics as key technologies. Biometric analysis technologies and methods are finding their way into all areas of life, changing people's daily lives. In particular the areas of sensor technology, biometric recognition process, and the general tendency toward convergence of information and communication technologies are stimulating the Big Data research. The conquest of mass markets through sensor and biometric recognition processes can sometimes be explained by the fact that mobile, web-based terminals are equipped with a large variety of different sensors. More and more users come this way into contact with the sensor technology or with the measurement of individual body characteristics. Due to the more stable and faster mobile networks, many people are permanently connected to the Internet using their mobile devices, providing connectivity an extra boost.

With the development of apps, application software for mobile devices such as smartphones (iPhone, Android, BlackBerry, Windows Phone) and Tablet computer, the application culture of biosurveillance changed significantly, since these apps are strongly influenced by the dynamics of the *bottom-up* participation. Andreas Albrechtslund speaks in this context of the "Participatory Surveillance" (2008) on the social networking sites, in which biosurveillance increasingly assumes itself as a place for open production of meaning and permanent negotiation, by providing comment functions, hypertext systems, and ranking and voting procedures through collective framing processes. This is the case of the sports app *Runtastic*, monitoring different sports activities, using GPS, mobile devices, and sensor technology, and making information, such as distance, time, speed, and burned calories, accessible and visible for friends and acquaintances in real

time. The *Eatery* app is used for weight control and requires its users the ability to do self-optimization through self-tracking. Considering that health apps also aim to influence the attitudes of their users, they can additionally be understood as persuasive media of *Health Governance*. With their feedback technologies, the apps facilitate not only issues related to healthy lifestyles but also multiply the social control over compliance with the health regulations in peer-to-peer networks. Taking into consideration the network connection of information technology equipment, as well as the commercial availability of biometric tools (e.g., “Nike Fuel,” “Fitbit,” “iWatch”) and infrastructure (apps), the biosurveillance is frequently associated, in the public debates, to dystopian ideas of a society of control biometrically organized.

Organizations and networks for health promotion, health information, and health education and formation observed with great interest that, every day, millions of users worldwide search for information about health using the Google search engine. During the influenza season, the searches for flu increase considerably, and the frequency of certain search terms can provide good indicators of flu activity. Back in 2006, Eysenbach evaluated in a study on “Infodemiology” or “Infoveillance” the Google *AdSense* click quotas, with which he analyzed the indicators of the spread of influenza and observed a positive correlation between increasing search engine entries and increased influenza activity. Further studies on the volume of search patterns have found that there is a significant correlation between the number of flu-related search queries and the number of people showing actual flu symptoms (Freyer-Dugas et al. 2012). This epidemiological correlation structure was subsequently extended to provide early warning of epidemics in cities, regions, and countries, in cooperation with the 2008 established *Google Flu Trends* in collaboration with the US authority for the surveillance of epidemics (CDC). On the *Google Flu Trends* website, users can visualize the development of influenza activity both geographically and chronologically. Some studies criticize that the predictions of the Google project are far above the actual flu cases.

Ginsberg et al. (2009) point out that in the case of an epidemic, it is not clear whether the search engines behavior of the public remains constant and thus whether the significance of *Google Flu Trends* is secured or not. They refer to the medialized presence of the epidemic as distorting cause of an “Epidemic of Fear” (Eysenbach 2006, p. 244), which can lead to miscalculations concerning the impending influenza activity. Subsequently, the prognostic reliability of the correlation between increasing search engine entries and increased influenza activity has been questioned. In recent publications on digital biosurveillance, communication processes in online networks are more intensely analyzed. Especially in the field of Twitter Research (Paul and Dredze 2011), researchers developed specific techniques and knowledge models for the study of future disease development and work backed up by context-oriented sentiment analysis and social network analysis to hold out the prospect of a socially and culturally differentiated biosurveillance.

Further Reading

- Albrechtslund, A. (2008). Online social networking as participatory surveillance. *First Monday*, 13(3). Online: <http://firstmonday.org/ojs/index.php/fm/article/viewArticle/2142/1949>.
- Brownstein, J. S., et al. (2009). Digital disease detection – Harnessing the web for public health surveillance. *The New England Journal of Medicine*, 360(21), 2153–2157.
- Burkom, H. S., et al. (2008). Decisions in biosurveillance tradeoffs driving policy and research. *Johns Hopkins technical digest*, 27(4), 299–311.
- Eysenbach, G. (2006). Infodemiology: Tracking flu-related searches on the Web for syndromic surveillance. In *AMIA Annual Symposium, Proceedings 8/2*, 244–248.
- Freyer-Dugas, A., et al. (2012). Google Flu Trends: Correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases*, 54(15), 463–469.
- Ginsberg, J., et al. (2009). Detecting influenza epidemics using search engine query data. In *Nature. International weekly journal of science* (Vol. 457, pp. 1012–1014).
- Paul, M. J., & Dredze, P. (2011). You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the Fifth International AAAI Conference on Weblogs*

and *Social Media*. Online: www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/.../3264.

Walters, R. A., et al. (2010). Data sources for biosurveillance. In G. Voeller John (Ed.), *Wiley handbook of science and technology for homeland security* (Vol. 4, pp. 2431–2447). Hoboken: Wiley.

Blockchain

Laurie A. Schintler

George Mason University, Fairfax, VA, USA

Overview of Blockchain Technology

Blockchain technology is one of the hallmarks of the Fourth Industrial Revolution (4IR). A blockchain is essentially a decentralized, distributed, and immutable ledger. The first significant application of blockchain technology was to bitcoin markets in the early 1990s. Since then, the uses of blockchain technology have expanded enormously, going well beyond bitcoin and cryptocurrency. Indeed, blockchain is now a pervasive technology in academia and government, and across and within industries and sectors. In this regard, some emerging applications include banking and financial payments and transfers, supply chain management, insurance, voting, energy management, retail trade, crowdfunding, public records, car leasing, cybersecurity, transportation, charity, scholarly communications, charity, government, health care, online music, real estate, criminal justice, and human resources.

Unlike centralized ledgers, blockchain records transactions between parties directly without third-party involvement. Each transaction is vetted and authenticated by powerful computer algorithms running across all the blocks and all the users. Such algorithms typically require consensus across the nodes, where different algorithms are used for this purpose. The decentralized distributed ledgers are updated asynchronously by adding a new block that is cryptographically “mined” based on a preset number of compiled

transactions. Typically, mining a block involves finding a solution to a cryptographic puzzle with varying levels of difficulty set by an algorithm. Each new block contains cryptographically hashed information on the most recent transactions and all previous transactions. Blocks are integrated in a chain-like manner, hence the name blockchain. All data on a blockchain is encrypted and hashed. Once validated and added to the blockchain, a transaction can never be tampered with or removed. In some cases, blockchain transactions are automated via “smart contracts,” which are agreements between two or more parties in the form of computer code; a transaction is only triggered if the agreed-upon conditions are met.

Blockchains are a way to establish trust in organizational (e.g., corporate) and personal transactions, which is fundamental to removing uncertainty. Several types of uncertainty face all transactions: (1) knowing the identity of the partners in a transaction, i.e., knowing whom one is dealing with; (2) transparency including the prehistory of conditions leading up to the transaction; and (3) recovering the loss associated with a transaction that fails. Establishing trust is how each of these uncertainties gets managed, which has traditionally been done through intermediaries that exact a cost for ensuring trust among the transacting parties. Since blockchain removes and replaces third parties, where no trust is required between those involved in transactions, it is characterized as a “trustless” system. (On the other hand, one can argue that the network of machines in a blockchain constitutes the intermediary, but one in a different guise than a conventional institutional third party.)

So, how specifically does one trust the identity of a transacting party? Each party is assigned a pair of cryptographically generated and connected electronic keys, a public key is known to the world, and a private key is known only to the party that owns it. These pairs of keys can be stored in the form of hard copies (paper copies) or relatively secure digital wallets owned by individuals. There are a number of public-key

cryptography-based algorithms. One of the most widely used kind of algorithm is known as the “Elliptic Curve Digital Signature Algorithm” (ECDSA).

Any transaction encrypted with a private key can only be decrypted by its associated public key and vice versa. For example, if Sender A wants to transact with Receiver B, then A encrypts the transaction with B’s public key and then signs it with the private key owned and known only to A. The receiving party B can verify the identity of A by using A’s public key and subsequently decrypt the transaction with his/her private key. There are many variations on how to use public-key cryptography for transactions, including transactions that involve multisignature protocol for transactions among multiple parties. In brief, public-key cryptography technology has enabled trustworthy peer-to-peer transactions among total strangers.

There are other ways in which blockchain promotes trust. First, it is generally a transparent system in which all blocks of information (history and changes) reside in a copy of the distributed ledger maintained by nodes (users). Second, with a blockchain system, there is no need for transacting parties to know each other as information about them and their transactions are known not only by each party but also by all users (parties). Therefore, the information is verifiable by all blockchain participants. Finally, blockchain provides mechanisms for recourse in the event of failed transactions. Specifically, recourse can be built into the block data or executed through smart contracts.

Blockchain and Big Data

Blockchain and big data is a “marriage made in heaven.” On the one hand, big data analytics are needed to vet the massive amounts of information added to a blockchain and arrive at a consensus regarding a transaction’s validity. On the other hand, blockchain provides a means for addressing the limitations of big data and the challenges associated with its use and application.

Indeed, big data is far from perfect. It tends to be fraught with noise, biases, redundancies, and other imperfections. Another set of issues relates to data provenance and data lineage, i.e., where the data comes from and how it has been used along the way. Once data is acquired, transmitted, and stored, such information should ideally be recorded for future use. However, with big data, this poses some difficulties. Big data tends to change hands frequently, where at each stop, it gets repurposed, repackaged, and reprocessed. Thus, the history of the data can get lost as it travels from one person, place, or organization to another. Moreover, in the case of proprietary and personally sensitive information, the data attributes tend to be hidden, complicating matters further. Finally, big data raises various kinds of privacy concerns. Many big data sources – including from transactions – contain sensitive, detailed, and revealing information about individuals, e.g., about their finances, personal behavior, or health and medical conditions. Such information may be intentionally or inadvertently exposed or used in ways that violate someone’s privacy. Data produced, consumed, stored, and transmitted in cyberspace are particularly vulnerable in these regards.

Blockchain technology can help to improve the quality, trustworthiness, traceability, transparency, privacy, and security of big data in several ways. As blockchains are immutable ledgers, unauthorized modification of any data added to them is virtually impossible. In other words, once the data is added to the blockchain, there is only a minimal chance that it can be deleted or modified. Data integrity and data security are further enhanced, given that transactions must be vetted and authenticated before being added to the blockchain. Additionally, since no form of personal identification is needed to initiate and use a blockchain, there is no central server with this information that could be compromised. Lastly, blockchain automatically creates a detailed and permanent record of all transactions (data) added to it, thus facilitating activities tied to data provenance and documentation.

The Dark Side of Blockchain

While blockchain does improve data security, it is not a completely secure system. More specifically, decentralized distributed tamper-proof blockchain technologies, although secure against most common cyber threats, can be vulnerable. For example, the blockchain mining operation can be susceptible to 51% attack where a party or a group of parties possess enough computing power to control the mining of new blocks. There is also a possibility of orphan blocks with legitimate transactions to be created, which never get integrated into the parent blockchain. Moreover, while blockchain technology is generally hack-proof due to its decentralized and distributed nature of operations, the rise of central exchanges for facilitating transactions across blockchains is open to cyberattacks. So are the digital wallets used by individuals to store their public/private keys.

Blockchain also raises privacy concerns. Indeed, blockchain technology is not entirely anonymous; instead, it is “pseudonymous,” where data points do not refer to any particular individuals, but multiple transactions by a single person can be combined and correlated to reveal their identity. This problem is compounded in public blockchains, which are open to many individuals and groups. The immutable nature of blockchain also makes it easy for anyone to “connect the dots” about individuals on the blockchain.

Finally, blockchain is touted as a democratizing technology where any person, organization, or place in the world can use and access it. However, in reality, specific skills and expertise, and enabling technologies (e.g., the Internet, broadband), are required to use and exploit blockchain. In this regard, the digital divides are a barrier to blockchain adoption for certain individuals and entities.

Considering all these issues and challenges, we need to develop technical strategies and public policy to ensure that all can benefit from blockchain while no one is negatively impacted or harmed by the technology.

Cross-References

- [Data Integrity](#)
- [Data Provenance](#)
- [Fourth Industrial Revolution](#)
- [Privacy](#)

Further Reading

- Deepa, N., Pham, Q. V., Nguyen, D. C., Bhattacharya, S., Gadekallu, T. R., Maddikunta, P. K. R., et al. (2020). A survey on blockchain for big data: Approaches, opportunities, and future directions. *arXiv preprint arXiv, 2009, 00858*.
- Karafiloski, E., & Mishev, A. (2017, July). Blockchain solutions for big data challenges: A literature review. In *IEEE EUROCON 2017-17th international conference on smart technologies* (pp. 763–768). IEEE.
- Nofer, M., Gomber, P., Hinz, O., & Schiereck, D. (2017). Blockchain. *Business & Information Systems Engineering*, 59(3), 183–187.
- Schintler, L. A., & Fischer, M. M. (2018). Big data and regional science: Opportunities, challenges, and directions for future research.
- Swan, M. (2015). *Blockchain: Blueprint for a new economy*. Sebastopol: O'Reilly Media.
- Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017, June). An overview of blockchain technology: Architecture, consensus, and future trends. In *2017 IEEE international congress on big data (BigData congress)* (pp. 557–564). IEEE.

Blogs

Ralf Spiller

Macromedia University, Munich, Germany

A blog or Web log is a publicly accessible diary or journal on a website in which at least one person, the blogger, posts issues, records results, or writes down thoughts. Often the core of a blog is a list of entries in chronological order. The blogger or publisher is responsible for the content, contributions are often written from a first-person perspective. A blog is for authors and reader an easy tool to cover all kinds of topics. Often, comments or discussions about

an article are permitted. Thus, blogs serve as a medium to gather, share, and discuss information, ideas, and experiences.

History

The first weblogs appeared in the mid-1990s. They were called online diaries and were websites on which Internet users periodically made entries about their own lives. From 1996, services such as Xanga have been set up that enabled Internet users to set up easily their own weblogs.

In 1997 one of the first blogs was started that still exists. It was called Scripting News, set up by Dave Winer. After a rather slow start, such sites grew rapidly from the late 1990. Xanga, for example, grew from 100 blogs in 1997 to 20 million in 2005. In recent years, blogging is also used for business in so-called corporate blogs. Also many news organizations like newspapers and TV-stations operate blogs to expand their audience and get feedback from readers and listeners.

According to Nielsen Social, a consumer research company, in 2006 there were about 36 million public blogs in existence, in 2009 about 127 million and in 2011 approximately 173 million. On September 2014, there were around 202 million Tumblr and more than 60 million WordPress blogs in existence worldwide. The total number of blogs can only be estimated but should be far more than 300 million in 2014.

Technical Aspects

Weblogs can be divided into two categories. First, the ones operated by a commercial provider allowing usage after a simple registration. Second, those that are operated by the respective owners on their individual server or webspace mostly under their own domain. Well-known providers of blog communities are Google's Blogger.com, WordPress, and Tumblr. Several social networks offer also blog functionalities to its members.

For the operation of an individual weblog on own web space, one needs at least a special weblog software and a rudimentary knowledge of HTML and the used server technology. Since blogs can be customized easily to specific needs, they are also often used as pure content management systems (CMS). Under certain circumstances, such websites are not perceived as blogs. From a pure technical point of view, all blogs are content management systems.

One of the important features of Web log software is online maintenance, which is performed through a browser-based interface (often called a dashboard), which allows users to create and update the contents of their blogs from any online browser. This software also supports the use of external client software to update content using an application programming interface. Web log software commonly includes plugins and other features that allow automatic content generation via RSS or other types of online feeds.

The entries, also called postings, blog posts or posts are the main components of a weblog. They are usually listed in reverse chronological order; the most recent posts can be found at the top of the weblog. Older posts are usually listed in archives. The consecutive posts on a specific topic within a blog are called a thread. Each entry, in some weblog systems, each comment has a unique and unchanging, permanent Web address (URL). Thus, other users or bloggers can directly link to the post. Web feeds, for example, rely on these permanent links (permalinks).

Most weblogs provide the possibility to leave a comment. Such a post is then displayed on the same page as the entry itself or as a popup. A web feed contains the contents of a weblog in a unified manner, and it can be subscribed via a feed reader. With this tool, the reader can take a look at several blogs at the same time and monitor new posts. There are several technical formats for feeds. The most common are RSS and Atom.

A blogroll is a list of other blogs or websites that a blogger endorses, commonly references or is affiliated with. A blogroll is generally found on one of the blog's side columns.

A weblog client (blog client) is an external client software that is used to update blog content

through an interface other than the typical Web-based version provided by blog software. There are desktop or mobile interfaces for blog posting. They provide additional features and capabilities, such as offline blog posting, better formatting, and cross-posting of content to multiple blogs.

Typology

Blogs can be segmented according to various criteria. From a content perspective, certain kinds of blogs are particularly popular: travel blogs, fashion blogs, technology blogs, corporate blogs, election blogs, warblogs, watch blogs, and blog novels. Other kinds of blogs are pure link blogs (annotated link collections), moblogs (Mobile Blogs), music blogs (MP3 blogs), audio blogs (podcasts), and vlogs (video blogs). Micro-blogging is another type of blogging, featuring very short posts like quotes, pictures, and links that might be of interest.

Blog lists like Blogrank.io provide useful information about the most popular blogs on a diverse range of topics. Several blog search engines are used to search blog contents, such as Bloglines.

The collective community of all blogs is known as the blogosphere. It has become an invaluable source for citizen journalism – that is, real time reporting about events and conditions in local areas that large news agencies or newspapers do not or cannot cover. Discussions in the blogosphere are frequently used by the media as a reflection of public opinion on various topics.

Characteristics

Empirical studies show that blogs emphasize personalization, audience participation in content creation, and story formats that are fragmented and interdependent with other websites.

Sharon Meraz (2011) shows that blogs exercise social influence and are able to weaken the influence of elite, traditional media as a singular power in issue interpretation within

networked political environments. Thus, blogs provide an alternative space to challenge the dominant public discourse. They are able to question mainstream representations and offer oppositional counter-discourses. Sometimes they are viewed as a form of civic, participatory journalism. Following this idea, they represent an extension of media freedom.

In certain topics like information technology, blogs challenge classic news websites and compete directly with their readers. This leads sometimes to innovations in the journalistic practice, for example, that online news sites adopt blog features.

Farrell and Drezner (2008) argue that under specific circumstances, blogs can socially construct an agenda or interpretive frame that acts as a focal point for mainstream media, shaping and constraining the larger political debate. This happens, for example, when key Web logs focus on a new or neglected issue.

Policy

Many human rights activists, especially in countries like Iran, Russia, or China, use blogs to publish reports on human rights violations, censorship, and current political and social issues without censorship by governments. Bloggers, for example, reported on the violent protests during the presidential elections in Iran in 2009 or the political upheavals in Egypt in 2012 and 2013. These blogs were an important source for news in Western media.

Blogs are harder to control than broadcast or even print media. As a result, totalitarian and authoritarian regimes often seek to suppress blogs and/or to punish those who maintain them.

Many politicians use blogs and similar tools like Twitter and Facebook particularly during election campaigns. US president Barack Obama was one of the first who used them effectively during his two presidential campaigns in 2009 and 2012. President Trump's tweets are carefully observed all over the world. Also nongovernmental organizations (NGOs) use blogs for their campaigns.

Blogs and Big Data

Big data usually refers to data sets defined by their volume, velocity, and variety. Volume refers to the magnitude of data, velocity to the rate at which data are generated and the speed at which it should be analyzed and acted upon and variety to the structural heterogeneity in a dataset.

Blogs are usually composed of unstructured data. This is the largest component of big data and is available as audio, images, video, and unstructured text. It is estimated that the analytics-ready structured data forms only a subset of big data of about 5% (Gandomi and Haider 2015).

Analyzing blog content implies dealing with imprecise data. This is a characteristic of big data, which is addressed by using tools and analytics developed for management and mining of uncertain data.

Performing business intelligence (BI) on blogs is quite challenging because of the vast amount of information and the lack of commonly adopted methodology for effectively collecting and analyzing such information. But the software is continually advancing and delivering useful results, for example, about product information in blogs.

According to Gandomi and Haider (2015), analytics of blogs can be classified into two groups: Content-based analytics and structure-based analytics. The first one focuses on data posted by users such as customer feedback and product reviews. Such content is often noisy and dynamic. The second one puts emphasis on the relationships among the participating entities. It is also called social network analytics. The structure of a social network is modeled through a framework of nodes and edges, representing participants and relationships. They can be visualized via social graphs and activity graphs.

Analytic tools can extract implicit communities within a network. One application is companies that try to develop more effective product recommendation systems. Social influence analysis evaluates the participants' influence, quantifies the strengths of connections, and uncovers the patterns of influence diffusion in a network. This information can be used for viral marketing to enhance brand awareness and adoption.

Various techniques can be used to extract information from the blogosphere: Blogs can be analyzed via sentiment analysis. Sentiment can vary by demographic group, news source, or geographic location. Results show opinion tendencies in popularity or market behavior and might also serve for forecasts regarding certain issues. Sentiment maps can identify geographical regions of favorable or adverse opinions for given entities.

Blogs can also be analyzed via content analysis methods. These are ways to gather rich, authentic, and unsolicited customer feedback. Information technology advances continuously and increasingly large numbers of blogs facilitate blog monitoring as a cost-effective method for service providers like hotels, restaurants, or theme parks.

Trends

Some experts see the emergence of Web logs and their proliferation as a new form of grassroots journalism. Mainstream media increasingly rely on information from blogs, and certain prominent bloggers can play a relevant role in the agenda setting process of news. In conclusion blogs have become together with other social media tools an irrefutable part of the new media eco systems with the Internet at its technical core.

Corporate blogs are used internally to enhance the communication and culture in a corporation or externally for marketing, branding, or public relations purposes. Some companies try to take advantage of the popularity of certain blogs and encourage these bloggers via free tests and other measures to post positive statements about products or services. Most bloggers do not see themselves as journalists and are open to cooperation with companies. Some blogs have become serious competitors for mainstream media since they are able to attract large readerships.

Cross-References

- [Content Management System \(CMS\)](#)
- [Sentiment Analysis](#)

Further Reading

- Blumenthal, M. M. (2005). Toward an open-source methodology. What we can learn from the blogosphere. *Public Opinion Quarterly*, 69(5, Special Issue), 655–669.
- Farrell, H., & Drezner, D. (2008). The power and politics of blogs. *Public Choice*, 134, 15–30.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management*, 35, 137–144.
- Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.
- Meraz, S. (2011). The fight for ‘how to think’: Traditional media, social networks, and issue interpretation. *Journalism*, 12(1), 107–127.

Border Control/Immigration

Btihad Ajana

King's College London, London, UK

Big Borders: Smart Control Through Big Data

Investments in the technologies of borders and in the securitization of movement continue to be one of the top priorities of governments across the globe. Over the last decade, there has been a notable increase in the deployment of various information systems and biometric solutions to control border crossing and fortify the digital as well as physical architecture of borders. More recently, there has been a growing interest in the techniques of big data analytics and in their capacity to enable advanced border surveillance and more informed decision-making and risk management. For instance, in Europe, programs such as Frontex and EUROSUR are examples of big data surveillance currently used to predict and monitor movements across EU borders. While in Australia, a recent big data system called Border Risk Identification System has been developed by IBM for the Australian Customs and Border Protection Service for the purpose of improving border management and targeting so-called “risky travellers.”

In this discussion, I argue that with big data come “big borders” through which the scope of control and monopoly over the freedom of movement can be intensified in ways that are bound to reinforce “the advantages of some and the disadvantages of others” (Bigo 2006, p. 57) and contribute to the enduring inequalities underpinning international circulation. Relatedly, I will outline some of the ethical issues pertaining to such developments.

To begin with, let us consider some of the definitions of big data. Generally, big data are often defined as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze” (McKinsey Global Institute 2011). They therefore require more enhanced technologies and advanced analytic capacities. Although emphasis is often placed on the “size” aspect, big data are by no means merely about large data. Instead they are more about the networked and relational aspect (Manovich 2011; Boyd and Crawford 2011). It is the power of connecting, creating/unlocking patterns, and visualizing correlations that makes big data such a seductive field of investment and enquiry for many sectors and organizations.

Big data can be aggregated from a variety of sources including web search histories, social media, online transactions records, mobile technologies and sensors that generate and gather information about location, and any other source where digital trails are left behind knowingly or unknowingly. The purpose of big data analytics is primarily about prediction and decision-making, focusing on “why events are happening, what will happen next, and how to optimize the enterprise’s future actions” (Parnell in Field Technologies Online 2013). In the context of border management, the use of big data engenders a “knowledge infrastructure” (Bollier 2010, p. 1) involving the aggregation, computation, and analysis of complex and large size contents which attempt to establish patterns and connections that can inform the process of deciding on border access, visa granting, and other immigration and asylum related issues. Such process is part and parcel of the wholesale automation of border securitization whereby border control is increasingly being conducted remotely, at a distance and well before

the traveller reaches the physical border (Broeders 2007; Bigo and Delmas-Marty 2011). More specifically, this involves, for instance, the use of Advance Passenger Records (APR) and information processing systems to enable information exchange and passenger monitoring from the time an intending passenger purchases an air ticket or applies for a visa (see for example the case of Australia's Advance Passenger Processing and the US Advance Passenger Information System). Under such arrangements, airlines are required to provide information on all passengers and crew, including transit travellers. This information is collected and transmitted to border agencies and authorities for processing and issuing passenger boarding directives to airlines prior to the arrival of aircrafts (Wilson and Weber 2008). A chief purpose of these systems is the improvement of risk management and securitization techniques through data collection and processing.

However, and as Bollier (2010, p. 14) argues, "more data collection doesn't mean more knowledge. It actually means much more confusion, false positives and so on." Big data and their analytical tools are, as such, some of the techniques that are being fast-tracked to enable more sophisticated ways of tracking the movement of perceived "risky" passengers. Systems such as the Australian Border Risk Identification System function through the scanning and analysis of massive amounts of data accumulated by border authorities over the years. They rely on advanced data mining techniques and analytical solutions to fine tune the knowledge produced out of data processing and act as a digital barrier for policing border movement and a tool for structuring intelligence, all with the aim to identify in advance suspected "high risk" passengers and facilitate the crossing of low risk ones. The argument is that automated big data surveillance systems make border control far more rigorous than what was previously possible. However, these data-driven surveillance systems raise a number of ethical concerns that warrant some reflection.

Firstly, there is the issue of *categorization*. Underlying border surveillance through big data is a process of sorting and classification, which

enables the systematic ordering, profiling, and categorization of the moving population body into pattern types and distinct categories. This process contributes to labeling some people as risky and others as legitimate travellers and demarcating the boundaries between them. In supporting the use of big data in borders and in the security field, Alan Bersin, from the US Department of Homeland Security, describes the profiling process in the following terms: "'high-risk' items and people are as 'needles in haystacks'. [Instead] of checking each piece of straw, [one] needs to 'make the haystack smaller', by separating low-risk traffic from high-risk goods or people" (in Goldberg 2013).

Through this rationality of control and categorization, there is the danger of augmenting the function of borders as spaces of "triage" whereby some identities are given the privilege of quick passage, whereas other identities are arrested (literally). The management of borders through big data technology is indeed very much about creating the means by which freedom of mobility can be enabled, smoothened, and facilitated for the qualified elite; the belonging citizens, all the while allowing the allocation of more time and effort for additional security checks to be exercised on those who are considered as "high risk" or "problematic" categories. The danger of such rationality and modality of control, as Lee (2013) points out, is that governments can target and "track undocumented migrants with an unheard of ease, prevent refugee flows from entering their countries, and track remittances and travel in ways that put migrants at new risks." The deployment of big data can thus become an immobilizing act of force that suppresses the movement of certain categories and restricts their access to spaces and services. With big data, the possibilities of control might be endless: governments might be able to

predict the next refugee wave by tracking purchases, money transfers and search terms prior to the last major wave. Or connect the locations of recipients of text messages and emails to construct an international network and identify people vulnerable to making the big move to join their family or spouse abroad. (If the NSA can do it, why not Frontex?) Or, an even more sinister possibility-

identify undocumented migrant clusters with greater accuracy than ever before by comparing identity and location data with government statistics on who is legally registered. (ibid.)

Another ethical concern relates to the issue of *projection*, which is at the heart of big data techniques. Much of big data analytics and the risk management culture within which it is embedded are based on acts of projection whereby the future itself is increasingly becoming the object of calculative technologies of simulation and speculative algorithmic probabilities. This techno-culture is based on the belief that one can create “a grammar of futur antérieur” by which the future can be read as a form of past in order to manage risk and prevent unwanted events (Bigo 2006). Big data promise to offer such grammar through their visualization techniques and predictive algorithms, through their correlations and causations. However, as Kerr and Earle (2013) argue, big data analytics raises concerns vis-à-vis its power to enable “a dangerous new philosophy of preemption,” one that operates by unduly making assumptions and forming views about others without even “encountering” them. In the context of border management and immigration control, this translates into acts of power, performed from the standpoint of governments and corporations, which result into the construction of “no-fly lists” and the prevention of activities that are perceived to generate risk, including the movement of potential asylum seekers and refugees.

What is at issue in this preemption philosophy is also a sense of reduced individual agency. The subjects of big data predictions are often unaware of the content and the scale of information generated about them. They are often unable to respond to or contest the “categorical assumptions” made about their behaviors and activities and the ensuing projections that affect many aspects of their lives, rights, and entitlements. Given the lack of transparency and the one-way character of big data surveillance, people are often kept unaware of the nature and extent of such surveillance and left without the chance to challenge the measures and policies that affect them in fundamental ways, such as criteria of access and so on. Autonomy and the ability to act in an informed and

meaningful way are significantly impaired, as a result. We are, as such, at risk of “being defined by algorithms we can’t control” (Lowe and Steenson 2013) as the management of life and the living becomes increasingly reliant on data and feedback loops. In this respect, one of the ethical challenges is certainly a matter of “setting boundaries around the kinds of institutional assumptions that can and cannot be made about people, particularly when important life chances and opportunities hang in the balance” (Kerr and Earle 2013). Circulation and movement are no exception.

The third and final point to raise here relates to the implications of big data on understandings and practices of *identity*. In risk management and profiling mechanisms, identity is “assumed to be anchored as a source of prediction and prevention” (Amoore 2006). With regard to immigration and border management, identity is indeed one of the primary targets of security technologies whether in terms of the use of biometrics to *fix* identity to the person’s “body” for the purpose of identification and identity authentication (Ajana 2013) or in terms of the deployment of big data analytics to construct predictive profiles to establish who might be a “risky” traveller. Very often, the identity that is produced by big data techniques is seen as disembodied and immaterial, and individuals as being reduced to bits and digits dispersed across a multitude of databases and networks and identified by their profiles rather than their subjectivities. The danger of such perception lies in the precluding of social and ethical considerations when addressing the implications of big data on identity, as individuals are seldom regarded in terms of their anthropological embeddedness and embodied nature. An embodied approach to the *materiality* of big data and identity is therefore needed to contest this presumed separation between data and their physical referent and the ever-increasing abstraction of people. This is crucial, especially when the identities at issue are those of vulnerable groups such as asylum seekers and refugees whose lives and potentialities are increasingly being caught up in the biopolitical machinery of bureaucratic institutions and their sovereign web of biopower.

Finally, it is hoped that this discussion has managed to raise awareness of some of the pertinent ethical issues concerning the use of big data in border management and to stimulate further debates on these issues. Although the focus of this discussion has been on the negative implications of big data, it is worth bearing in mind that big data technology also carries the potential to benefit vulnerable groups if deployed with an ethics of care and in the spirit of helping migrants and refugees as opposed to controlling them. For instance, and as Lee (2013) argues, big data can provide migration scholars and activists with more accurate statistics and help them fight back against “fear-mongering false statistics in the media,” while enabling new ways of understanding the flows of migration and enhancing humanitarian processes. As such, conducting further research on the empowering and resistance-enabling aspects of big data is certainly worth pursuing.

Further Reading

- Ajana, B. (2013). *Governing through biometrics: The biopolitics of identity*. Basingstoke: Palgrave Macmillan.
- Amoore, L. (2006). Biometric borders: Governing mobilities in the war on terror. *Political Geography*, 25, 336–351.
- Bigo, D. (2006). Security, exception, ban and surveillance. In D. Lyon (Ed.), *Theorising surveillance: The panopticon and beyond*. Devon: Willan Publishing.
- Bigo, D., & Delmas-Marty, M. (2011). The state and surveillance: Fear and control. http://cle.ens-lyon.fr/anglais/the-state-and-surveillance-fear-and-control-131675.kjsp?RH=CDL_ANG100100#P4.
- Bollier, D. (2010). The promise and peril of big data. http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.
- Boyd, D., & Crawford, K. (2011). Six provocations for big data. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431.
- Broeders, D. (2007). The new digital borders of Europe: EU databases and the surveillance of irregular migrants. *International Sociology*, 22(1), 71–92.
- Field Technologies Online. (2013). Big data: Datalogic predicts growth in advanced data collection as business analytics systems drive need for more data and innovation. <http://www.fieldtechnologiesonline.com/doc/big-data-datalogic-data-collection-systems-data-innovation-0001>.
- Goldberg, H. (2013). Homeland Security official gives lecture on borders and big data. <http://www.michigandaily.com/news/ford-school-homeland-security-lecture>.
- Kerr, I., & Earle, J. (2013). Prediction, preemption, presumption: How big data threatens big picture privacy. <http://www.stanfordlawreview.org/online/privacy-and-big-data/prediction-preemption-presumption>.
- Lee, C. (2013). Big data and migration – What’s in store? <http://noncitizensoftheworld.blogspot.co.uk/>.
- Lowe, J., & Steenson, M. (2013). The new nature vs. nurture: Big data & identity. http://schedule.sxsw.com/2013/events/event_IAP5064.
- Manovich, L. (2011). Trending: The promises and the challenges of big social data. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.
- McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
- Wilson, D., & Weber, L. (2008). Surveillance, risk and preemption on the Australian border. *Surveillance and Society*, 5(2), 124–141.

Brain Research Through Advancing Innovative Neurotechnologies

► White House BRAIN Initiative

Brand Monitoring

Chiara Valentini

Department of Management, Aarhus University,
School of Business and Social Sciences, Aarhus,
Denmark

Introduction

Brand monitoring is the act of searching and collecting large datasets on brands with the purpose of evaluating brand performance and value as perceived by consumers and the public in general. Today, a lot of data on brands is collected online. Online brand monitoring is about scanning, gathering, and analyzing content that is published on the Web. Online data is machine-readable, has explicitly defined meanings, and is

linked to other external datasets (Bizer 2009). Given that most data is today complex and unstructured and requires different storage and processing, brand monitoring often relies on big data analytics which consists of collecting, organizing, and analyzing large and diverse datasets from various databases. Brand monitoring is a central activity for the strategic brand management of an organization and any organized entity. A brand is an identifier of a product, a service, an organization or even a person's main characteristics and qualities. Its main scope is to differentiate products, services or an individual's qualities from those of competitors through the use of a specific name, term, sign, symbol, or design, or a combination of them. In marketing, Kotler (2000) defines a brand as a "name associated with one or more items in the product line that is used to identify the source of character of the item(s)" (p. 396). Brands have existed for centuries, yet, the modern understanding of a brand as something related to trademarks, attractive packaging, etc., that signify a guarantee of product or service authenticity, is a phenomenon of late nineteenth century (Fullerton 1988).

Brands and Consumers

The brand concept became popular in marketing discipline as a company tactic to help customers and consumers to identify specific products or services from competitors but also to communicate their intangible qualities (Kapferer 1997). Today, anything can be branded, for example, an organization, a person, and a discipline. The concept of branding, that is, the act of publicizing a product, service, organization, person, etc., through the use of a specific brand name, has become more than a differentiation tactic. It has turned into a strategic management discipline. The scope is to create a strong product, service, organization, and personal identity that can lead to positive images and attitudes among existing and potential consumers, customers and even the general public.

Reflecting on the impact of branding in business organizations Kapferer (1997) noted a shift

in consumers' interests from desiring a specific commodity to desiring a precise, branded type of good or service. He observed that certain brands represent something more than a product or service; they own a special place in the minds of consumers. Companies are, indeed, trying to gain a special place in the minds of consumers by focusing on creating brand values and charging consumers who purchase those brands extra dollars for these specific values. Brand values can be of functional nature, that is, they have specific characteristics related to product or service quality. Brand values can also be of symbolic nature, that is, they can possess intangible characteristics such as particular meanings resulting from holding and owning specific brands, as well as from the act of brand consumption.

Brand monitoring is an important step for evaluating brand performance and brand values and, in general, for managing a brand.

Understanding Consumption Motives and Practices

Kotler (2000) argues that the most important function of marketers is to create, maintain, protect, and enhance brands. Along the same line of thoughts, Tuominen (1999) postulates that the real capital of companies is their brands and the perception of these brands in the minds of potential buyers. Because brands have become so important for organizations, the studying of strategic brand management, integrated marketing communication, and consumer behavior has become more and more important in marketing research. Due to the tangible and intangible nature of brand values, an important area of study in brand management is the identification and assessment of how changing elements of the marketing mix impact customer attitudes and behaviors. Another important area of study is understanding consumption motives and practices. Diverse dataset analytics have become important tools for the study of both these areas. In explaining how consumers consume, Holt (1995) identifies four typologies of consumption

practices: *consuming as experience*, *consuming as integration*, *consuming as classification*, and *consuming as a play*. Consuming as experience represents the act of consuming a product or service because consuming it provokes some enjoyment, a positive experience on its own. Consuming as integration is the act of consuming with the scope of transferring the meanings that specific brands have into own identity. It is an act that serves the purpose of constructing a personal identity. Consumers can use brands to strengthen a specific social identification and use consumption as practice for making themselves be recognizable for the objects they own and use. This act is consuming as classification. Finally, consuming as a play reflects the idea that through consumption people can develop relationships and relate to others. This last purpose of consuming was later used to explain a specific type of brand value, called the *linking value* (Cova 1997). The linking value is often referred to a product or service's contribution to establishing or reinforcing bonds between individuals.

These four typologies of consumption practices show that in developed economies people buy and consume products not only for their basic human needs and the functional values of products, but often for their symbolic meanings (Sassatelli 2007). Symbolic meanings are not created in a vacuum, but are often trends and tendencies coming from different social and cultural phenomena. They are, in the words of McCracken (1986), borrowed from the culturally constituted world, which is a world where meanings about objects are shaped and changed by a diverse variety of people. For example, companies such as Dior have for long employed cultural icons, such as the French actress Catherine Deneuve, in the Chanel No. 5 perfume campaigns, in their product advertisements to transpose cultural meanings of such icons into the product brand value. Research indicates that these iconic associations have a strong attitudinal effect on consumers, since consumers think about brands as if they were celebrities or famous historical figures (Aaker 1997). Studies on brand association emerged as well as those related to brand awareness. In marketing it is well

recognized that established brands reduce marketing costs because they increase the brand visibility and thus help in getting consumer consideration, provide reasons to buy, attract new customers via awareness and reassurance, and create positive attitudes and feelings that can lead to brand locality (Aaker 1991). Therefore, brands can produce equities.

Brand Monitoring for Marketing Research

Besides identifying the benefits that brands can provide to organizations, marketing research has been interested in investigating how to measure brand value to explain its contribution to organizations' business objectives. This value is measured through brand equity. A brand equity is "a set of brand assets and liabilities linked to a brand, its name and symbol, that adds to or subtracts from the value provided by a product or service to a firm and/or to that firm's customer" (Aaker, 1991, p. 15). Aaker's brand equity comprises four dimensions: brand loyalty, brand awareness, brand associations, and perceived quality. Positive brand equity is obtained through targeted marketing communications activities that generate "customer awareness of the brand and the customer holds strong, unique, and favorable brand associations in memory" (Christodoulides and de Chernatony 2004, p.169). Traditionally brand equity was measured through indicators such as price premium, customer satisfaction or loyalty, perceived quality, brand leadership or popularity, perceived brand value, brand personality, organizational associations, brand awareness, market share, market price, and distribution coverage. Yet with the increased popularity of Internet and social media, a lot of organizations are using online channels to manage their brands and the values they can offer to their customers and consumers. Christodoulides and de Chernatony (2004) propose to include other brand equity indicators to assess online brand value. These are online brand experience, interactivity, customization, relevance, site design, customer service, order fulfillment, quality of brand

relationships, communities, as well as website logs statistics.

Conclusion

Brand monitoring helps organizations to collect large sets of data on online brand experiences which encompasses all points of interaction between the customer and the brand in the virtual space. Online experiences are also about the level of interactivity, quality of the site, or social media design and customization that organizations can offer to their online customers. Through the use of specific software computing big data analytics, organizations can collect large and diverse datasets on individuals' preferences and they can systematically analyze and interpret them to provide unique content of direct relevance to each customer. Companies like Amazon track their customers' purchases and provide a customized list of suggested items when customers revisit their company websites (Ansari and Mela 2003). Other brand equity indicators are Web log metrics, the number of hits, the number of revisits and view time per page, number of likes, shares, and retweets (Christodoulides and de Chernatony 2004). Information on viewers is collected via web bugs. A Web bug (also known as a tracking bug, pixel tag, Web beacon, or clear gif) is a graphic in a website or a graphic-enabled e-mail message. Sentiment analysis also known as opinion mining is another type of social media analytics that allows companies to monitor the status of consumers and publics' opinions on their brands (Stieglitz et al. 2014). Behavioral analytics is another approach to collect and analyze large scale datasets on consumers or simply web visitors' behaviors. According to the Privacy Rights Clearinghouse (2014, October), companies regularly engage in behavioral analytics with the purpose of monitoring individuals, their web searches, the visited pages, the viewed content, their interactions on social networking sites, and the products and services they purchase.

Brand monitoring is an important component of brand equity measurement, and today there exists a number of software applications and

even sites that allow companies to monitor and assess the status of their brands. For instance, Google offers *Google Trends* and *Google Analytics*; these are tools that monitor search traffic of a company and its brand. Integrated monitoring services such as *Hootsuite* for monitoring Twitter, Facebook, LinkedIn, WordPress, Foursquare and Google+ conversations in real-time and *SocialMention* for seeking web for user-generated content, such as blogs, comments, bookmarks, events, news, videos, and microblogging services, etc., have also been used for collecting specific brand dataset content. Social media companies collect large strings of social data on a regular basis, sometimes for company related purposes other times for selling to other companies that seek information on their brands on those social networking sites (boyd and Crawford 2012). When they do not buy such datasets, organizations can acquire information on consumers' opinions on their brand experience, satisfaction, and overall impression on their brands by simply scanning and monitoring online conversations in social media, Internet fora and sites. Angry customers, dissatisfied employees, and consumer activists use the Web and social media as weapons to attack brands, organizations, political figures and celebrities. Therefore, it has become paramount for any organization and prominent individual to have in place mechanisms to gather, analyze and interpret big data on people's opinions on their brands. Yet as boyd and Crawford (2012) pointed out, there are still issues on relying only on large datasets from Web sources as these are often unreliable, not necessarily objective or accurate. Big data analytics are often taken out of context, and this means that datasets lose meaning and value, especially when organizations are seeking to assess their online brand equity. Furthermore, ethical concerns about anonymity and privacy of individuals can emerge when organizations collect datasets online.

Cross-References

- [Behavioral Analytics](#)
- [Business Intelligence Analytics](#)

- Facebook
- Google Analytics
- Online Advertising
- Privacy
- Sentiment Analysis

Further Reading

- Aaker, J. L. (1991). *Managing brand equity*. New York: The Free Press.
- Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research*, 34(3), 347–356.
- Ansari, A., & Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, 40(2), 131–145.
- Bizer, C. (2009). The emerging web of linked data. *Intelligent Systems*, 24(5), 87–92.
- boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662–679.
- Christodoulides, G., & de Chernatony, L. (2004). Dimensionalising on- and offline brands' composite equity. *Journal of Product & Brand Management*, 13(3), 168–179.
- Cova, B. (1997). Community and consumption: Towards a definition of the linking value of products or services. *European Journal of Marketing*, 31(3/4), 297–316.
- Fullerton, R. A. (1988). How modern is modern marketing? Marketing's evolution and the myth of the production era. *Journal of Marketing*, 52(1), 108–125.
- Holt, D. B. (1995). How consumers consume: A typology of consumption practices. *Journal of Consumer Research*, 22(1), 1–16.
- Kapferer, J.-N. (1997). *Strategic brand management*. London, UK: Kogan Page.
- Kotler, P. (2000). *Marketing management. The millennium edition*. Upper Saddle River: Prentice Hall.
- McCracken, G. (1986). Culture and consumption: A theoretical account of the structure and movement of the cultural meaning of consumer goods. *Journal of Consumer Research*, 13(1), 71–84.
- Privacy Rights Clearinghouse. (2014, October). *Fact sheet 18: Online privacy: Using the internet safely*. <https://www.privacyrights.org/online-privacy-using-internet-safely>. Accessed on 7 Nov 2014.
- Sassatelli, R. (2007). *Consumer culture: History, theory and politics*. London, UK: Sage.
- Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, 6(2), 89–96.
- Tuominen, P. (1999). Managing brand equity. *Liiketaloudellinen Aikakauskirja – The Finnish Journal of Business Economics*, 48(1), 65–100.

Business

Magdalena Bielenia-Grajewska
Division of Maritime Economy, Department of
Maritime Transport and Seaborne Trade,
University of Gdansk, Gdansk, Poland
Intercultural Communication and
Neurolinguistics Laboratory, Department of
Translation Studies, University of Gdansk,
Gdansk, Poland

Business consists of the “profit seeking activities and enterprises that provide goods and services necessary to an economic system” (Boone and Kurtz 2013, p. 5) and can be understood from different perspectives. Taking a macroapproach, business can be treated as the sum of activities directed at gaining money or any other profit. The meso-point of investigation concentrates on business as a sector or a type of industry. Examples may include, among others, automotive business, agricultural business, and media business. As in the case of other types, the appearance of new enterprises is connected with different factors. For example, the growing role of digital technologies has resulted in the advent of different types of industries, such as e-commerce or e-medicine. Applying a more microperspective, a business is an organization, a company, or a firm, which focuses on offering goods or services to customers. In addition, business can refer to one's profession or type of conducted work. No matter which perspective is taken, business is a complex entity being influenced by different factors and, at the same time, influencing other entities in various ways. Moreover, irrespective of level of analysis, big data plays an increasingly central role in business intelligence, management, and analysis, such that basic studies and understandings of business rely on big data as a determinant feature.

Subdomains in Business Studies

Different aspects of business have been incorporated as various subdomains or subdisciplines by

which it has been studied. For example, a broad subdomain is *marketing*, focusing on how to make customers interested in products or services and facilitating their selection. Researchers and marketers are interested in, e.g., branding and advertising. Branding is connected with making customers aware that a brand exists, whereas advertising is related to using verbal and nonverbal tools as well as different channels of communication to make products visible on the market and then selected by users.

Another subdomain is *corporate social responsibility* (CSR), referring to ethical dimensions of corporate performance, with particular attention to social and environmental issues, since modern companies pay more and more attention to ethics. One of the reasons that modern companies are increasingly concerned with ethics is market competitiveness; firms observe the ways they are viewed by different stakeholders since many customers, prior to selecting products or services, might consider how a company respects the environment, takes care of local communities, or supports social initiatives. CSR can be categorized according to different notions. For example, audience is the basis for many CSR activities, analyzed through the prism of general stakeholders, customers, local communities, or workers. CSR also can be studied by examining the types of CSR activities focused on by a company (Bielenia-Grajewska 2014). Thus, the categorization can include strengthening the potential of workers and taking care of the broadly understood environment. The internal dimension is connected with paying attention to the rights and wishes of workers, whereas the external one deals with the needs and expectations of stakeholders located outside the company.

Since modern organizations are aimed at achieving particular goals and targets, *management* stands out as a subdomain of contemporary business. It can be classified as the set of tools and strategies for reaching corporate aims by using different types of capital and resources. It is the role of managers to make available human and nonhuman assets in order to achieve company goals. Regarding the human aspect of management, there are different profiles of managers,

depending on the type of a company, its organizational structure, and organizational culture. In some organizations, especially large ones and corporations, there are top (or senior), middle, and first-line managers who perform different functions. In small business, the structure is less complex, often confined to a single manager responsible for different jobs. Moreover, there are different characteristics that define a good manager. For example, he/she must be a good leader; leadership skills include the possession of social power to influence or persuade people to achieve certain goals. Studies on leadership have stressed issues such as inborn features or learned expertise (soft and occupational skills), which make a person a good leader. It should also be stated that leadership styles, although sharing some common features, may be culture-specific. Thus, no leadership style applies to all situations and communities; the way a group of people is to be guided depends on the features, expectations, needs, and preferences of a given community. In the management literature, the most often discussed types of leadership are autocratic (authoritarian) leadership, democratic leadership, and laissez-faire (free-rein) leadership. Autocratic leaders do not discuss their visions and ideas with subordinates but implement their decisions without any prior consultancy with workers. Democratic leaders, on the other hand, allow for the active role of employees in decision-making processes. Thus, this way of leadership involves mutual cooperation in the process of making and implementing decisions. Laissez-faire leadership makes the workers responsible for decision-making. The role of a supervisor is mainly to monitor and communicate with employees as far as their decisions are concerned. As provided in the definition, management is connected with organizing human capital. Thus, the way individuals work and contribute to the performance of an organization determines the way a company is organized.

As a concept, *human resources* denote the set of tools and strategies used to optimize the performance of workers in a company. The staff employed by a department of human resources is responsible for, among other things, recruiting, employing, and laying off workers, along with organizing vocational

training and other methods of improving the professional skills of personnel. Thus, such concepts as human capital and talent management are used to denote the skills and abilities of employees within the area of human resources.

A prominent subdomain of business studies is *finance*, since running a business is intimately related to managing its financial sphere. Finance is connected with assets, liabilities, and equities of a company. For example, accounting (or financial reporting) in companies focuses on keeping a financial record of corporate activities, paying taxes, and providing information on the financial situation of companies for interested stakeholders, an area that is increasingly complex and growing with the incorporation of big data analytics.

Yet another domain of study important for contemporary business is *business law*, comprising the set of regulations connected with corporate performance. It concerns the legal sphere of creation, production, and sale of products. It should also be stressed that the various spheres do not only determine contemporary business as such but also influence themselves. For example, the financial sphere is determined by business law, whereas management depends on available human resources in a company and its crucial characteristics. Thus, contemporary business is not only shaped in terms of the subdomains but also shapes them.

Factors Shaping Contemporary Business

An important factor shaping the way business functions is its environment, broadly conceived. Today's competitive business world is affected by the competitive environment, the global environment, the technological environment, and the economic environment (Pride et al. 2012). Analysis of contemporary businesses from different perspectives allows enumeration of at least seven crucial determinants of contemporary business, as indicated in Fig. 1.

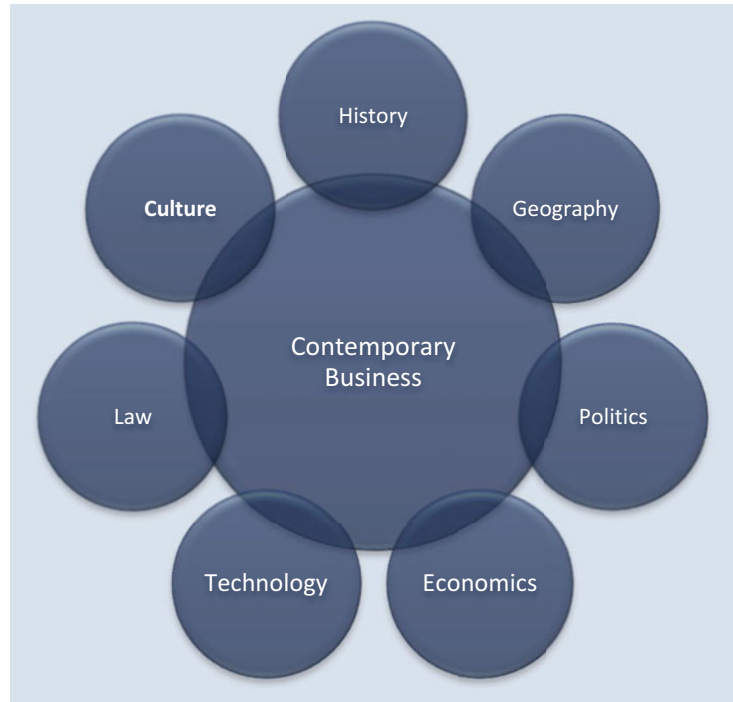
First, *history* determines the performance of companies since the way business entities function mirrors the history not only of a company but also of a state. In addition, the traces of history are observed in the way firms are managed, being

represented in, e.g., how types of management are related to the past political systems in a given state. As far as branding products is concerned, history is used to stress the long tradition or experience of a given company to produce merchandise or deliver services. Thus, a company operating for a long time on the market is often regarded as trustworthy and experienced in a given industry.

Also important is *geography* since location determines contemporary business in different ways. For one thing, it shapes the profile of a company since a given business has to adjust to available geographical conditions. For example, there are fisheries located near the water reservoirs such as seas, oceans, or lakes. In that case, setting up a business near a given geographical location limits the cost of transportation. In other situations, geographical characteristics may serve as a barrier in running a given type of company due to the lack of required resources or limited access to its premises for most customers. Moreover, a particular geographical location may work as a tool of branding a given company. For example, mountains or villages may be associated with fresh air and nature, and thus, products offered by the companies operating in such areas are associated with healthy food.

Another factor is *politics* since decisions made by politicians influence the way companies function. For example, the type of governing can enhance or diminish the amount of people interested in running private companies. *Economics* is another determinant of contemporary business. From a macroeconomic perspective, such notions as production, consumption, inflation, growth, and unemployment influence the way contemporary businesses function. For example, low unemployment may result in companies having to increase wages or salaries. The microeconomic consideration of households, sellers, or buyers shapes the way companies adjust prices and the level of production. Also, *technology* is a factor; the twenty-first century can be characterized as the age of extensive technological developments in all spheres of life. For example, modern companies rely on the Internet in advertising their products and communicating with stakeholders. In that

Business, Fig. 1 Main determinants of contemporary business



way, technology is responsible for lowering the costs of communication and making it more effective in comparison with standard methods of exchanging information, such as face-to face interactions or regular mail. It should also be stressed that companies do not exist in a vacuum and their performance is determined by the legal sphere, i.e., the *law*. Laws and other legal regulations shape the way a company is set up, run, and closed.

Finally, *culture* is probably the most complex factor of all the determinants. It can be understood in many ways, with the perspective of constituents being presented as the first approach. Communication is one of the most important and most visible representations of culture in modern business, understood by taking inner and outer dimensions into account. The division of internal (with and among workers) and external (with broadly understood stakeholders) communication can be applied. Also online and offline communications are a main determinant of interaction. Online communication involves all types of

interactions taking place in the web. It is connected with sending e-mails, using social media networking tools, such as Facebook or Twitter, posting information at websites, etc. Off-line communication is understood as all types of communicative exchanges that do not involve the Internet. Thus, this notion entails direct and indirect interaction without the online exchange of data.

Communication can also be discussed in terms of its elements. The main typology involves verbal and nonverbal tools of communication. As far as the verbal sphere is concerned, language is responsible for shaping corporate linguistic policies, determining the role of corporate lingo as well as the national language of a country a company operates in and the usage of professional languages and dialects. In addition, it also shapes the linguistic sphere of those going abroad, e.g., expatriates who also have to face linguistic challenges in a new country. Moreover, language is also not only the means of corporate communication, but it is also a sphere of activity that is regulated in

companies. Corporate linguistic rights and corporate linguistic capital are issues that should be handled with care by managers since properly managed *company linguistic identity capital* offers possibilities to create a friendly and efficient professional environment for both external and internal communication (Bielenia-Grajewska 2013a). As far as the linguistic aspect is concerned, related tools can be divided into literal and nonliteral ones. Nonliteral tools encompass symbolic language, represented by, e.g., metonymies and metaphors. Taking metaphors into consideration, they turn out to be efficient in intercultural communication, relying on broadly understood symbols, irrespective of one's country of origin. Using a well-known domain to describe a novel one is a very effective strategy of introducing new products or services on the market. Paying attention to available connotations, metaphors often make stakeholders attracted to the merchandise. In addition, metaphors can be used to categorize types of contemporary business. For example, a modern company may be perceived as a teacher; introducing novel technologies and advancements in a given field results in customers having access to the latest achievements and products.

On the other hand, taking the CSR perspective into account, companies may become a paragon of protecting the environment, teaching the local community how to take care of their neighborhood. In addition, companies may promote a given lifestyle, such as eating healthy food, exercising more, or spending free time in an educational way. Another organizational metaphor is a learner, with companies playing the role of students in the process of organizational learning since they observe the performance of other companies and the behavior of customers. Apart from the verbal sphere, a company communicates itself through pictorial, olfactory, and audio channel. The pictorial corporate dimension is represented by, e.g., logos or symbols, whereas the audio one is visible in songs used in advertising or corporate jingles. The olfactory level is connected with all the smells related in some way to the company itself (e.g., the scent used in corporate offices) or its products.

Communication can also be divided by discussing the type of stakeholders participating in interaction. Taking into account the participants, communication can be classified as internal and external. *Internal corporate communication* entails interactions taking place between employees of a company. This dimension incorporates different types of discourse among workers themselves as well as between workers and employees, including such notions as hierarchy, power distance, and organizational communication policy. On the other hand, *external corporate communication* focuses on interactions with the broadly understood stakeholders. It involves communication with customers, local community, mass media, administration, etc. Communication is also strongly influenced by cultural notions, such as the type of culture shared or not shared by interlocutors. Apart from the discussed element-approach, one may also discuss culture as a unique set of norms and rules shared by a given community (be it ethnic, national, professional, etc.).

Another approach for investigating culture and its role for contemporary business is by looking at cultural differences. No matter which typology is taken into account, modern companies are often viewed through similarities and differences in terms of values they praise. The differences may be connected with national dichotomies or organizational varieties. For example, such notions as the approach to power or hierarchy in a given national culture are taken as a factor determining the way a given company is run. In addition, companies are also viewed through the prism of differences related to organizational values and leadership styles. It should be mentioned, however, that contemporary business is not only an entity influenced by different factors, but it is also an entity that influences others. Contemporary business influences the environment at both individual and societal level. Applying the microposition, modern companies construct the life of individuals. Starting with the tangible sphere, they shape the type of competence and skills required from the individuals, being the reason why people decide to educate or upgrade their qualification. They are often

the reasons why people decide to migrate or reorganize their private life. Taking the meso-level into account, companies determine the performance of other companies, through such notions as competitiveness, providing necessary resources, etc. The macrodimension is related to the way companies influence the state.

Types of Contemporary Business

One way of classifying businesses is by taking into account the type of ownership (Pride et al. 2012). This classification encompasses different types of individual and social ownership. For example, as in the case of *sole proprietorship*, one can run his or her own business, without employing any workers. When an individual runs a business with another person, it is called *partnership* (e.g., limited liability partnership, general partnership). On the other hand, corporations are companies or groups of companies acting as a single legal entity, having rights and liabilities toward their workers, shareholders, and other stakeholders. There is also a state (public) ownership when a business is run by the state. Another type of running a business is franchising. A franchisee is given the right to use the brand and marketing strategies of a given company to run its own store, restaurant, service point, etc.

Modern companies also can be studied from the perspective of profit and nonprofit statuses. Contemporary business can also be divided by taking into account different types of business and looking through the prism of how they are run and managed. In that case, leadership styles as well as corporate cultures can be examined. Contemporary business can also be discussed by analyzing its scope of activities. Thus, division into regional, national, and international companies can be used. Business also can be classified by analyzing type of industry. With the advent of new technologies and the growing popularity of the Internet, companies can also be subcategorized according to online-off-line distinctions. Thus, apart from standard types of business, nowadays more and more customers

opt for e-commerce, offering goods or services on the web.

Big Data and Researching Contemporary Business

The growing expectations of customers who are faced with multitudes of goods and services have led to the emergence of cross-domain studies that contribute to a complex picture of how stakeholders' expectations can be met. Thus, many researchers opt for multidisciplinary methods. A popular approach to investigating contemporary business is a network perspective. Network studies offer a multidimensional picture of the investigated phenomenon, drawing attention to different elements that shape the overall picture. Analyzing the application of network approaches in contemporary business, e.g., Actor-Network-Theory, stresses the role of living and nonliving entities in determining corporate performance. It may be used to show how not only individuals but also, e.g., technology, mobile phones, and office artifacts influence the operational aspects of contemporary business (Bielenia-Grajewska 2011). The selection of methods is connected to the object of analysis. Thus, researchers use approaches such as observation to investigate corporate culture, interviews to focus on hierarchy issues, or expert panels to study management styles. Moreover, contemporary business can be researched by applying qualitative or quantitative approaches. The first are focused on researching a carefully selected group of individuals, factors, and notions in order to observe some tendencies, whereas the quantitative studies are related to dealing with relatively high numbers of people or concepts. Moreover, there are types of research, generally associated with other fields, which can provide novel data on modern companies. Taking linguistics as an example, discourse studies can be used to research how the selection of words and phrases influences the attitude of clients toward offered products. One of the popular domains in contemporary business is neuroscience. Its growing role in different modern disciplines has influenced, e.g., management, and resulted in

such areas of study as international neurobusiness, international neurostrategy, neuromarketing, neuroentrepreneurship, and neuroethics (Bielenia-Grajewska 2013b).

No matter which approach is taken into account, both researchers studying contemporary business and managers running the companies have to deal with an enormous amount of data of different types and sources (Roxburgh 2019). They can include demographic data (names, addresses, sex, ethnicity, etc.), financial data (income, expenditures), retail data (shopping habits), and data connected with transportation, education, health, and social media use. Sources of information include public, health, social security, and retail repositories as well as the Internet. The variety and volume of data and its complex features lead to many techniques that can be used by organizations to deal with these types of data. These can include techniques such as data mining, text mining, web mining, graph mining, network analysis, machine learning, deep learning, neural networks generic algorithms, spatial analysis, and search-based application (Olszak 2020).

Cross-References

- [Ethics](#)
- [Human Resources](#)
- [International Nongovernmental Organizations \(INGOs\)](#)
- [Semiotics](#)

Further Reading

- Bielenia-Grajewska, M. (2011). A potential application of actor network theory in organizational studies: The company as an ecosystem and its power relations from the ANT perspective. In A. Tatnall (Ed.), *Actor-network theory and technology innovation: Advancement and new concepts*. Hershey: Information Science Reference.
- Bielenia-Grajewska, M. (2013a). Corporate linguistic rights through the prism of company linguistic identity capital. In C. Akrivopoulou & N. Garipidis (Eds.), *Digital democracy and the impact of technology on governance and politics. New globalized perspectives*. Hershey: IGI Global.

- Bielenia-Grajewska, M. (2013b). International Neuromanagement. In D. Tsang, H. H. Kazeroony, & G. Ellis (Eds.), *The Routledge companion to international management education*. Abingdon: Routledge.
- Bielenia-Grajewska, M. (2014). CSR online communication: The metaphorical dimension of CSR discourse in the food industry. In R. Tench, W. Sun, & B. Jones (Eds.), *Communicating corporate social responsibility: Perspectives and practice* (Critical studies on corporate responsibility, governance and sustainability) (Vol. 6). Bingley: Emerald Group Publishing.
- Boone, L. E., & Kurtz, D. L. (2013). *Contemporary business*. Hoboken: Wiley.
- Celina, M. O. (2020). *Business intelligence and big data: Drivers of organizational success*. Boca Raton: CRC Press.
- Olszak, C. (2020). *Business Intelligence and Big Data: Drivers of Organizational Success*. Abingdon: CRC Press.
- Pride, W., Hughes, R., & Kapoor, J. (2012). *Business*. Mason: South Western.
- Roxburgh, E. (2019). *Business and big data: Influencing consumers*. New York NJ: Lucent Press.

Business Intelligence

- [Business Intelligence Analytics](#)

Business Intelligence Analytics

Feras A. Batarseh
College of Science, George Mason University,
Fairfax, VA, USA

Synonyms

[Advanced analytics](#); [Big data](#); [Business intelligence](#); [Data analytics](#); [Data mining](#); [Data science](#); [Data visualizations](#); [Predictive analytics](#)

Definition

Business Intelligence Analytics is a wide set of solutions that could directly and indirectly influence the decision making process of a business

organization. Many vendors build Business Intelligence (BI) platforms that aim to plan, organize, share, and present data at a company, hospital, bank, airport, federal agency, university, or any other type of organization. BI is the business umbrella that has analytics and big data under it.

Introduction

Nowadays, business organizations need to carefully gauge markets and take key decisions, quicker than ever before. Certain decision can steer the direction of an organization, and halt its progress, while other decisions can improve its place in the market and even increase profits. If BI is broken into categories, three organizational areas would emerge: *technological intelligence* (understanding the data, advancing the technologies used, and the technical footprint), *market intelligence* (studying the market, predicting where it is heading and how to react to its many variables), and *strategic intelligence* (dictates how

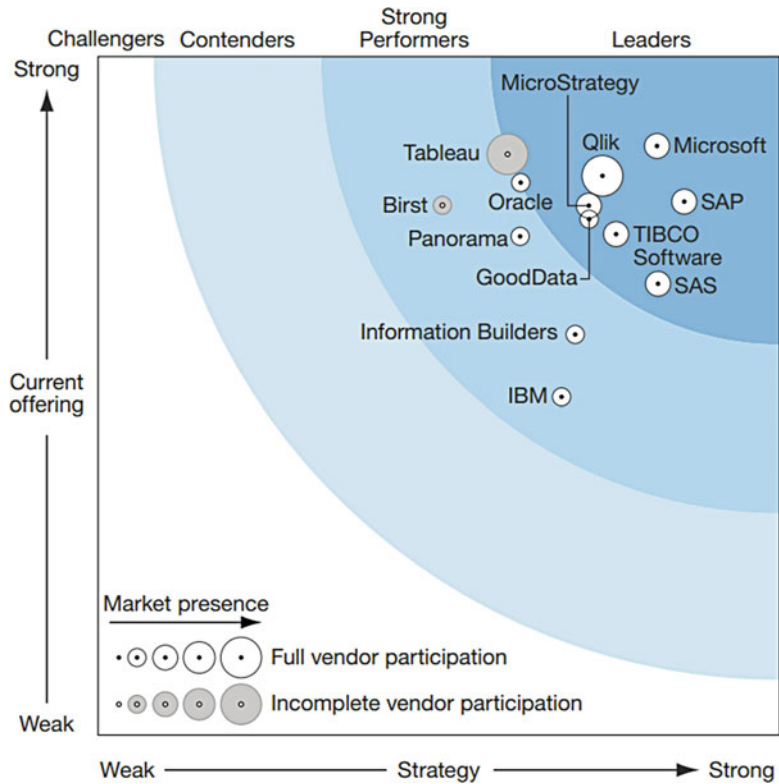
to organize, employ, and structure an organization from the inside; and how strategies affect the direction of an organization in general).

Main BI Features and Vendors

The capabilities of BI include decision support, statistical analysis, forecasting, and data mining. Such capabilities are achieved through a wide array of features that a BI vendor should inject into their software offering. Most BI vendors provide such features, however, the leading global BI vendors are: IBM (Watson Analytics), Tibco (Spotfire), Tableau, MicroStrategy, SAS, SAP (Lumira), and Oracle and Microsoft (PowerBI). Figure 1 below illustrates BI market leaders based on the power of execution and performance and the clarity of vision.

BI vendors provide a wide array of software, data, and technical features for their customers; the most commonplace features include database management, data organization and

Business Intelligence Analytics, Fig. 1 Leading BI vendors (Forrester 2015)



augmentation, data cleaning and filtering, data normalization and ordering, data mining and statistical analysis, data visualization, and interactive dashboards (Batarseh et al. 2017).

Many industries (such as healthcare, finance, athletics, government, education, and the media) adopted analytical models within their organizations. Although data mining research has been of interest to many academic researchers around the world for a long time, data analytics (a form of BI) did not see much light until it was adopted by the industry. Many software vendors (SAS, SPSS, Tableau, Microsoft, and Pentaho) shifted the focus of their software development to include a form of BI analytics, big data, statistical modeling, and data visualization.

BI Applications

As it is mentioned previously, BI has been deployed at many domains, some famous and successful BI applications include: (1) healthcare records collection and analysis, (2) predictive analytics for the stock market, (3) airport passengers flow and management analytics, (4) federal government decision and

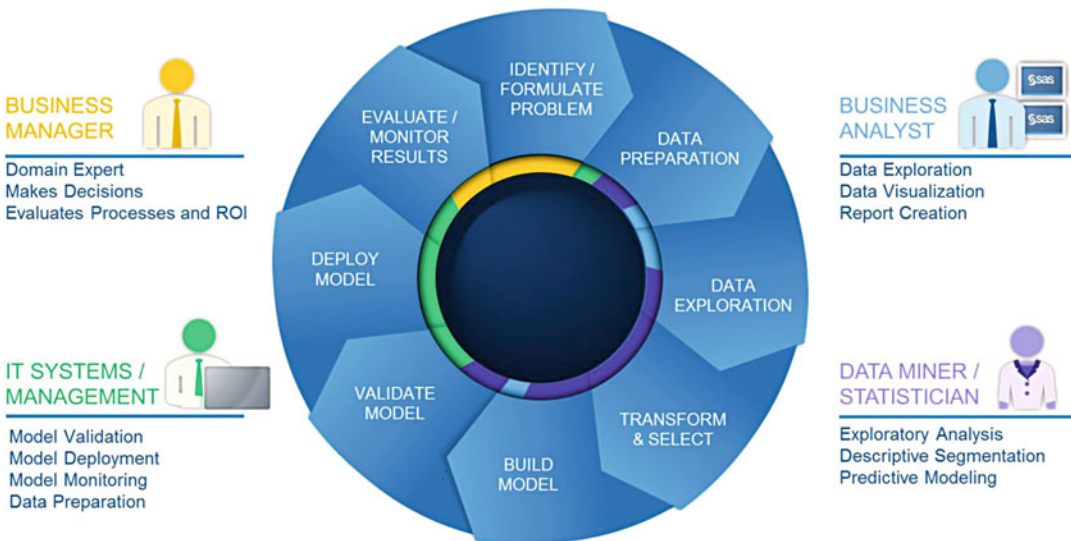
policy making, (5) geography, remote sensing, and weather forecasting, and (6) defense and army operations, among many other successful applications.

However, to achieve such decision-making support functions, BI relies heavily on structured data. Obtaining structured data is quite challenging in many cases, and data are usually raw, unstructured, and unorganized. Business organizations have data in forms of emails, documents, surveys, sheets, tables, and even meeting notes; furthermore, they have data for customers that can be aggregated at many different levels (such as weekly, monthly, or yearly), but to achieve successful applications, most organizations need to have a well-defined BI lifecycle. The BI lifecycle is introduced in the next section.

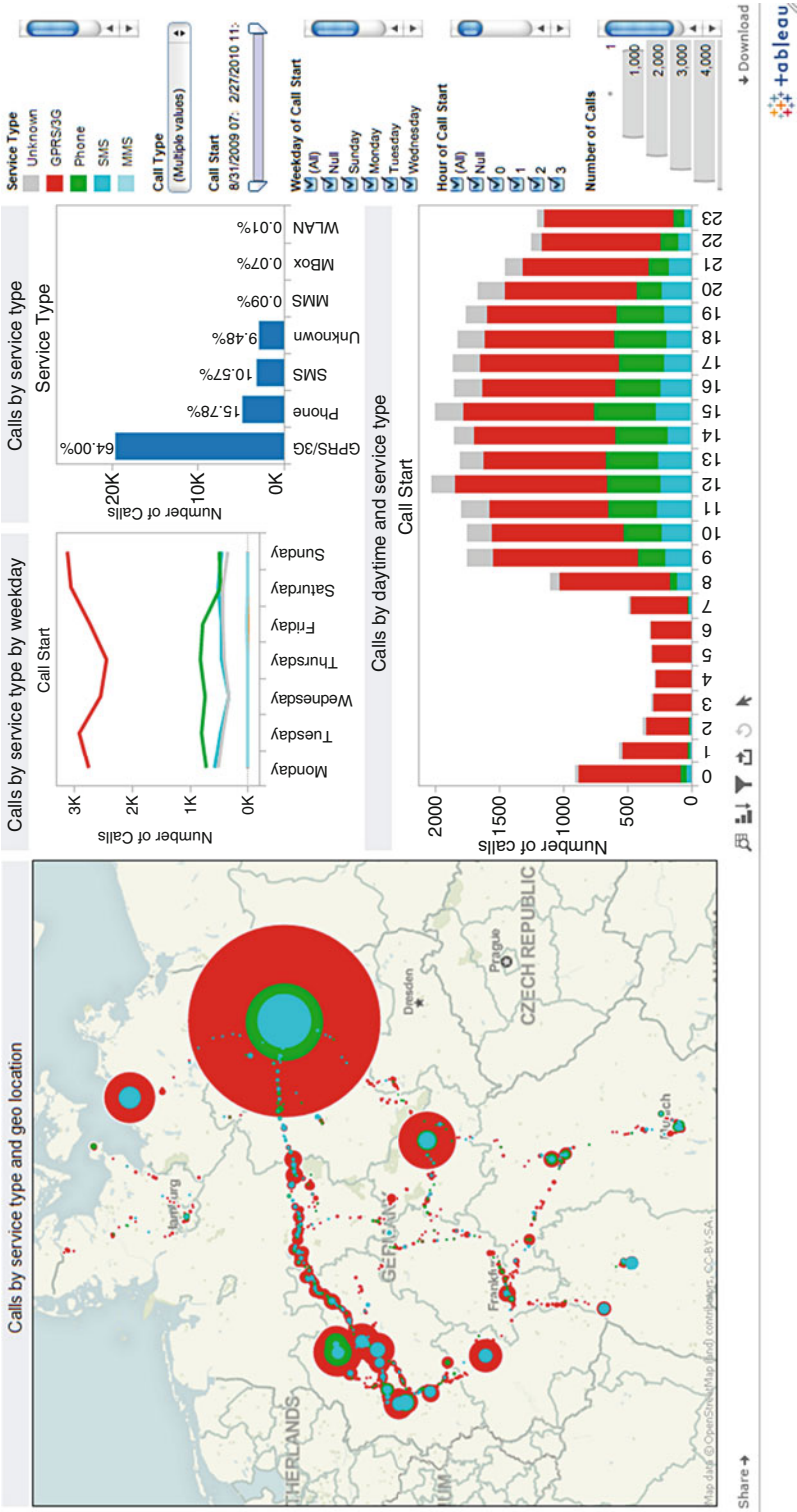
The BI Development Lifecycle

Based on multiple long and challenging deployments in many fields, trials, and errors, and many consulting exchanges with customers from a variety of domains, BI vendors coined a data management lifecycle model for BI. SAS provided that model (illustrated in Fig. 2).

THE PREDICTIVE ANALYTICS LIFECYCLE



Business Intelligence Analytics, Fig. 2 BI lifecycle model (SAS 2017)



Business Intelligence Analytics, Fig. 3 A dashboard (Tableau 2017)

The model includes the following steps: identify and formulate the problem; prepare the data (pivoting and data cleansing), data exploration (through summary statistics charts), data transformation and selection (select ranges, and create subsets), statistical model development (data mining), validation, verification and deployment; evaluate and monitor results of models; deliver the best model; and observe the results and refine (Batarseh et al. 2017). The main goal of the BI lifecycle is to allow BI engineers to transform the *big data* into useful reports, graphs, tables, and dashboards. Dashboards and interactive visualizations are the main outputs of most BI tools. Figure 3 shows an example output – a Tableau Dashboard (Tableau 2017).

BI outputs are usually presented on top of a data warehouse. The data warehouse is the main repository of all data that are created, collected, generated, organized, and owned by the organization. Data warehouses can host databases (such as Oracle databases), or big data that is unstructured (but organized through tools such as Hadoop). Each of the mentioned technologies has become essential in the lifecycle of BI and its outputs.

Different vendors have different weaknesses and strengths, most of which are presented in many market analysis studies presented in publications from McKinsey & Company, Gartner, and Forrester (Forrester 2015).

Conclusion

Business Intelligence (analytics built on top of data, in many cases big data) is a rapidly growing field of research and development and has attracted interest from academia and government but *mostly* from industry. BI analytics depend on many other software technologies and research areas of study such as data mining, machine learning, statistical analysis, art, user interfaces, market intelligence, artificial intelligence, and big data. BI has been used in many domains, and it is still witnessing a growing demand with many new applications. BI is a highly relevant and a very interesting area of study that is worth investing in at all venues and exploring at all levels.

Further Reading

- Batarseh, F., Yang, R., & Deng, L. (2017). *A comprehensive model for management and validation of federal big data analytical systems*. Published at Springer's journal of Big Data Analytics.
- Evelson, B. (2015). *The Forrester wave: Agile business intelligence platforms*. A report published by Forrester Research, Inc.
- SAS website and reports. (2017). Available at: http://www.sas.com/en_us/home.html.
- Tableau website and dashboards. (2017). Available at: <http://www.tableau.com/>.

Business-to-Community (B2C)

Yulia A. Levites Strekalova
College of Journalism and Communications,
University of Florida, Gainesville, FL, USA

Organizations of all types and sizes reap the benefits of web interactivity and actively seek to engage with their customers. Social communities and online engagement functionality offered by the web allow businesses to foster the creation of communities of customers and benefit from multi-way conversations and information exchanges that happen in these communities. Moreover, big data drives web development and the creation of online communities such that principles of business-to-community (B2C) engagement have applications beyond traditional marketing efforts and create value for associations, nonprofits, and civic groups. Online communities and social networks facilitate the creation, organization, and sharing of knowledge. As social organisms, communities go through life cycles of development and maturity and show different characteristics at different stages of development. Interactive communication and engagement techniques in the enterprise promise to have profound and far-reaching effects on how organizations reach and support customers as communities. Community building, therefore, requires a long-term commitment and ongoing efforts from organizations. Overall, communities may become an integral part of a business's operations and become an

asset if strategically engaged, or a liability if mismanaged.

Traditional business-to-consumer interactions relied on one-way conversations between businesses and their consumers. These conversations could take a form of interviews or focus groups as part of a market research initiative or a customer satisfaction survey. These methods are still effective in collecting customer feedback, but they are limited in their scope as they are usually guided by predefined research questions. As such, extended data collection activities force customers to provide information of interest to an organization rather than collect unaided information on the areas or products of specific interest or concern to the customers. Conversely, in a community communication environment, community members can pose questions themselves indicating what interests or concerns them most.

Community Development

Community planning requires organizations to establish a general strategy for community development, make decisions about the type of leadership in the community, decide on desired cultural characteristics of the knowledge, define the level of formality of the community management, decide on the content authorship, and develop a set of metrics to establish and measure the outcomes of community engagement.

The decision to maintain an online community creates an asset with a lot of benefits for the organizations, but it also creates a potential liability. Ill-maintained, unmoderated community communication may create negative company and brand perceptions among potential customers and prompt them to consider a different brand. Similarly, existing customers who do not get prompt responses in online community may not develop a strong tie to the company and its product and failing to engage with the product to the extent they could. These situations may have a long-term effect on the future product choices for the existing customers.

Rich Millington identifies three models facilitated by big data for building large active online

communities: the host-created model, the audience-created model, and the co-creation model. The host-created model of community building is a top-down approach, where an organization builds a community for its target audience and encourages its customers to actively participate in the online knowledge exchange. This approach relies on the organization's staff rather than volunteers to keep the community in the community active and offers the most control to the organization. It also requires the most control, strategic planning, and ongoing effort to grow and maintain the community. The audience-created model is a bottom-up approach, which is driven by the consumers themselves based on shared hobbies and fandom. In this case, organizations support and cultivate passion for their product through a loyal group of advocates. While this approach may be more cost-effective for organizations, its outcomes are also a lot less predictable and may be hard to measure. The last, co-creation, model is a hybrid of the first two, where an organization may provide technical support and a platform for communication and exchange of ideas, but individuals are drawn to the community to satisfy their information needs through group interaction and knowledge exchange.

Community Life Cycle

Margaret Brooks and her colleagues, discussing business-to-business (B2B) social communities, describe a four-stage community life cycle model. This four-stage model describes communities and their development as an *onboarding-established-mature-mitotic* continuum. Onboarding communities are new and forming. These communities usually attract early adopters and champions who want to learn more about a product or an organization. Members can contribute to the development of the community and a virtual collaborative space. At this stage, new community members are interested in supporting the community and creating benefits mutually shared between them and an organization. Onboarding communities are most vulnerable and need to

create interest among active community members to gain momentum and attract additional followers. Established communities are characterized by an established membership with leaders, advocates and followers. Members of these communities start to create informal networks and share knowledge among them. Mature communities, which have existed for a few years, are said to be virtually self-sustaining. These networks feature internal teams, area experts, and topical collaborations. The goal of mature communities is not to increase their membership but to keep existing members engaged in active communication and information exchange. Finally, mitotic communities are compared to a mother cell that grows and separates into daughter cells with identical characteristics. The separation could be based on a regional level for large communities or based on a development of different product lines. This process could be an indication that a community lost its core focus. Yet, it could also be a natural process that will lead to the creation of new established and mature communities with narrower focus.

Active Communities

Interaction functionality afforded by the web creates new opportunities for organizations to contact and connect with their customers by providing rich data and informational support and interactive communication through wikis and blogs. The use of product champions and community experts allows organizations to assist with problem resolution through discussion groups and online forums. Web-based modes of customer support creates opportunities to service tens of thousands of customers with just hundreds of employees and ongoing support with community expert volunteers. Additionally, customer-to-customer advocacy for an organization or its products may be more persuasive and powerful than organization's own marketing and advertising efforts.

Engaged communities create competitive advantage for companies that succeed in forming and supporting them and provide

valuable insights for companies. If organizations can build systems to collect and analyze data on consumer insights, community communication can feed ideas for new product development from the perspective of those who will be using the products. Communities can also drive marketing efforts by creating and spreading the buzz about new products and services. Here, the necessary ingredient for effective viral marketing is understanding of the audience that will support a new product and tailoring of communication to generate interest in this audience. Finally, communities can self-support themselves in real time. This support can provide a robust, powerful and sustaining environment for training and education and act as a product demonstration ground for potential new customers.

Active engagement is key to successful online community efforts and several strategies can help in increasing participation of the members. Lack of participation in an online community may be associated with several factors. For example, community groups are too segmented, which makes it hard for the existing and new community members to find the right group to ask a question. One indicator of this problem may be a low number of members per group or a low number of new questions posted to a discussion group.

Content relevance is another issue that may contribute to the community inactivity. The analysis of audience interests and the questions that audiences look to resolve through online community participation may not overlap with the company's immediate priorities, yet if content is not of value to the audience, large quantities of irrelevant content will not lead to large quantities of community engagement.

Ongoing engagement is the third area that may contribute to the lack of online participation. The rates of new members visting and registering in the community, the amount of information each member views in a session, participation in discussions, initiation of new discussions and other online communication behavior are all areas that may require an intervention. Combination of all these points may create audience profiles and

facilitate further audience analysis and comparison of individual community members against exiting and desired models of member engagement.

Numerous studies have shown that only 5–20% of online community members engage in active communication while the rest are passive information consumers. The attitudes of the latter, dominant group is harder to assess, yet the passive consumption of information does not mean the lack of engagement with the company's products. This does mean, however, that a few active online community members may act as opinion leaders, thus having a strong effect on the rest of the customers participating in an online community created largely through big data processes.

Cross-References

- ▶ [Cluster Analysis](#)
- ▶ [Profiling](#)
- ▶ [Sentiment Analysis](#)
- ▶ [Social Media](#)

Further Reading

- Brooks, M. (2013). *Developing B2B social communities: keys to growth, innovation, and customer loyalty*. CA: CA Technologies Press.
- Millington, R. (2014). *Buzzing communities: how to build bigger, better, and more active online communities*. Lexington: FeverBee.
- Simon, P. (2013). *Too big to ignore: the business case for big data*. Hoboken: Wiley.

Cancer

Christine Skubisz

Department of Communication Studies, Emerson College, Boston, MA, USA

Department of Behavioral Health and Nutrition, University of Delaware, Newark, DE, USA

Cancer is an umbrella term that encompasses more than 100 unique diseases related to the uncontrolled growth of cells in the human body. Cancer is not completely understood by scientists, but it is generally accepted to be caused by both internal genetic factors and external environmental factors. The US National Cancer Institute describes cancer on a continuum, with points of significance that include prevention, early detection, diagnosis, treatment, survivorship, and end-of-life care. This continuum provides a framework for research priorities. Cancer prevention includes lifestyle interventions such as tobacco control, diet, physical activity, and immunization. Detection includes screening tests that identify atypical cells. Diagnosis and treatment involves informed decision making, the development of new treatments and diagnostic tests, and outcomes research. Finally, end-of-life care includes palliative treatment decisions and social support. Large data sets can be used to uncover patterns, view trends, and examine associations between variables. Searching, aggregating, and cross-referencing large data sets is beneficial at all

stages within the cancer continuum. Sources of data include laboratory investigations, feasibility studies, clinical trials, cancer registries, and patient medical records. The paragraphs that follow describe current practices and future directions for cancer-related research in the era of big data.

Cancer Prevention and Early Detection

Epidemiology is the study of the causes and patterns of human diseases. Aggregated data allows epidemiologists to study why and how cancer forms. Researchers study the causes of cancer and ultimately make recommendations about how to prevent cancer. Data provides medical practitioners with information about populations at risk. This can facilitate proactive and preventive action. Data is used by expert groups including the American Cancer Society and the United States Preventive Services Task Force to write recommendations about screening for detection. Screening tests, including mammography and colonoscopy, have advantages and disadvantages. Evidence-based results, from large representative samples, can be used to recommend screening for those who will gain the largest benefit and sustain the fewest harms. Data can be used to identify where public health education and resources should be disseminated.

At the individual level, aggregated information can guide lifestyle choices. With the help of

technology, people have the ability to quickly and easily measure many aspects of their daily lives. Gary Wolf and Kevin Kelly coined this rapid accumulation of personal data the quantified self movement. Individual-level data can be collected through wearable devices, activity trackers, and smartphone applications. The data that is accumulated is valuable for cancer prevention and early detection. Individuals can track their physical activity and diet over time. These wearable devices and applications also allow individuals to become involved in cancer research. Individuals can play a direct role in research by contributing genetic data and information about their health. Health care providers and researchers can view genetic and activity data to understand the connections between health behaviors and outcomes.

Diagnosis and Treatment

Aggregated data that has been collected over long periods of time has made a significant contribution to research on the diagnosis and treatment of cancer. The Human Genome Project, completed in 2003, was one of the first research endeavors to harness large data sets. Researchers have used information from the Human Genome Project to develop new medicines that can target genetic changes or drivers of cancer growth. The ability to sequence the DNA of large numbers of tumors has allowed researchers to model the genetic changes underlying certain cancers.

Genetic data is stored in biobanks, repositories in which samples of human DNA are stored for testing and analysis. Researchers draw from these samples and analyze genetic variation to observe differences in the genetic material of someone with a specific disease compared to a healthy individual. Biobanks are run by hospitals, research organizations, universities, or other medical centers. Many biobanks do not meet the needs of researchers due to an insufficient number of samples. The burgeoning ability to aggregate data across biobanks, within the United States and internationally, is invaluable and has the potential to lead to new discoveries in the future.

Data is also being used to predict which medications may be good candidates to move forward into clinical research trials. Clinical trials are scientific studies that are designed to determine if new treatments and diagnostic procedures are safe and effective. Margaret Mooney and Musa Mayer estimate that only 3% of adult cancer patients participate in clinical trials. Much of what is known about cancer treatment is based on data from this small segment of the larger population. Data from patients who do not participate in clinical trials exists, but this data is unconnected and stored in paper and in electronic medical records. New techniques in big data aggregation have the potential to facilitate patient recruitment for clinical trials. Thousands of studies are in progress worldwide at any given point in time. The traditional, manual, process of matching patients with appropriate trials is both time consuming and inefficient. Big data approaches can allow for the integration of medical records and clinical trial data from across multiple organizations. This aggregation can facilitate the identification of patients for inclusion in an appropriate clinical trial. Nicholas LaRusso writes that IBM's supercomputer Watson will soon be used to match cancer patients with clinical trials. Patient data can be mined for lifestyle factors and genetic factors. This can allow for faster identification of participants that meet inclusion criteria. Watson, and other supercomputers, can shorten the patient identification process considerably, matching patients in seconds. This has the potential to increase enrollment in clinical trials and ultimately advance cancer research.

Health care providers' access to large data sets can improve patient care. When making a diagnosis, providers can access information from patients exhibiting similar symptoms, lifestyle choices, and demographics to form more accurate conclusions. Aggregated data can also improve a patient's treatment plan and reduce the costs of conducting unnecessary tests. Knowing a patient's prognosis helps a provider decide how aggressively to treat cancer and what steps to take after treatment. If aggregate data from large and diverse groups of patients were available in a

single database, providers would be better equipped to predict long-term outcomes for patients. Aggregate data can help providers select the best treatment plan for each patient, based on the experiences of similar patients. This can also allow providers to uncover patterns to improve care. Providers can also compare their patient outcomes to outcomes of their peers. Harlan Krumholz, a professor at the Yale School of Medicine, argued that the best way to study cancer is to learn from everyone who has cancer.

Survivorship and End-of-Life Care

Cancer survivors face physical, psychological, social, and financial difficulties after treatment and for the remaining years of their lives. As science advances, people are surviving cancer and living in remission. A comprehensive database on cancer survivorship could be used to develop, test, and maintain patient navigation systems to facilitate optimal care for cancer survivors.

Treating or curing cancer is not always possible. Health care providers typically base patient assessments on past experiences and the best data available for a given condition. Aggregate data can be used to create algorithms to model the severity of illness and predict outcomes. This can assist doctors and families who are making decisions about end-of-life care. Detailed information, based on a large number of cases, can allow for more informed decision making. For example, if a provider is able to tell a patient's family with confidence that it is extremely unlikely that the patient will survive, even with radical treatment, this eases the discussion about palliative care.

Challenges and Limitations

The ability to search, aggregate, and cross-reference large data sets has a number of advantages in the prevention and treatment of cancer. Yet, there are multiple challenges and limitations to the use of big data in this domain. First, we are limited to the data that is available. The data set

will always be incomplete and will fail to cover the entire population. Data from diverse sources will vary in quality. Self-reported survey data will appear alongside data from randomized, clinical trials. Second, the major barrier to using big data for diagnosis and treatment is the task of integrating information from diverse sources. Allen Lichter explained that 1.6 million Americans are diagnosed with cancer every year, but in more than 95% of cases, details of their treatments are in paper medical records, file drawers, or electronic systems that are not connected to each other. Often, the systems in which useful information is currently stored cannot be easily integrated. The American Association of Clinical Oncology is working to overcome this barrier and has developed software that can accept information from multiple formats of electronic health records. A prototype system has collected 100,000 breast cancer records from 27 oncology groups. Third, traditional laboratory research is necessary to understand the context and meaning of the information that comes from the analysis of big data. Large data sets allow researchers to explore correlations or relationships between variables of interest. Danah Boyd and Kate Crawford point out that data are often reduced to what can fit into a mathematical model. Taken out of context, results lose meaning and value. The experimental designs of clinical trials will ultimately allow researchers to show causation and identify variables that cause cancer. Bigger data, in this case more data, is not always better. Fourth, patient privacy and security of information must be prioritized at all levels. Patients are, and will continue to be, concerned with how genetic and medical profiles are secured and who will have access to their personal information.

Cross-References

- ▶ [Evidence-Based Medicine](#)
- ▶ [Health Care Delivery](#)
- ▶ [Nutrition](#)
- ▶ [Prevention](#)
- ▶ [Treatment](#)

Further Reading

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *Journal of the American Medical Association*, 309(13), 1351–1352.

Cell Phone Data

Ryan S. Eanes

Department of Business Management,
Washington College, Chestertown, MD, USA

Cell phones have been around since the early 1980s, but their popularity and ubiquity expanded dramatically at the turn of the twenty-first century as prices fell and coverage improved; this demand for mobile phones has steadily grown worldwide. The introduction of the iPhone and subsequent smartphones in 2007, however, drove dramatic change within the underlying functionality of cellular networks, given these devices' data bandwidth requirements for optimal function. Prior to the advent of the digital GSM (Global System for Mobile Communications, originally *Groupe Spéciale Mobile*) and CDMA (code division multiple access) networks in Europe and North America in the 1990s, cell phones solely utilized analog radio-based technologies. While modern cellular networks still use radio signals for the transmission of information, the content of these transmissions has changed to packets of digital data. Indeed, the amount of data generated, transmitted, and received by cell phones is tremendous, given that virtually every cell phone handset sold today utilizes digital transmission technologies. Furthermore, cellular systems continue to be upgraded to handle this enormous (and growing) volume of data.

This switch to digitally driven systems represents both significant improvements in data bandwidth and speeds for users as well as potential new products, services, and areas of new research based on the data generated by our devices. Therefore, this article will first outline in general terms the mechanics of cell phone data transmission and

reception, and the ways in which this data volume is managed via cellular networks. Consideration will also be given to consumer-facing industry practices related to cell data access, including data pricing and roaming charges. The article will conclude with a brief examination of the types of information that can be collected on cell phone users who access and utilize cellular data services.

Cell Phone Data Transmission and Reception

Digital cell phones still rely on radio technology for reception and transmission, just as analog cell phones do. However, digital information, which simply consist of 1's and 0's, is much more easily compressed than analog content; in other words, more digital "stuff" can fit into the same radio transmission that might otherwise only carry a solitary analog message. This switch to digital has meant that the specific radio bandwidths reserved for cell phone transmissions can now handle many more messages than analog technologies could previously.

When a call is placed using a digital cell phone on certain systems, including the AT&T and T-Mobile GSM networks in the United States, the cell phone's onboard digital-to-analog converter, or DAC, converts the analog sound wave into a digital signal, which is then compressed and transmitted via radio frequency (RF) to the closest cell phone tower, with GSM calls in the USA utilizing the 850 MHz and 1.9 GHz bands. As mentioned, multiple cell phone signals can occupy the same radio frequency thanks to time division multiple access, or TDMA. For example, say that three digital cell phone users are placing calls simultaneously and have been assigned to the same radio frequency. A TDMA system will break the radio wave into three sequential timeslots that repeat in succession, and bits of each user's signal will be assigned and transmitted in the proper slot. The TDMA technique is often combined with wide-band transmission techniques and "frequency hopping," or rapidly switching between available frequencies, in order to minimize interference.

Other types of digital networks, such as the CDMA (“code division multiple access”) system employed by Verizon, Sprint, and most other American carriers, take an entirely different approach. Calls originate in the same manner: A caller’s voice is converted to a digital signal by the phone’s onboard DAC. However, outgoing data packets generated by the DAC are tagged with a unique identifying code, and these small packets are transmitted over a number of the available frequencies available to the phone. In the USA, CDMA transmissions occur on the 800 MHz and 1.9 GHz frequency bands; each of these bands consists of a number of possible frequencies (e.g., the specific frequency 806.2 MHz is part of the larger 800 MHz band). Thus, a CDMA call’s packets might be transmitted on a number of frequencies simultaneously, such as 806.2 MHz, 808.8 MHz, 811.0 MHz, and so forth, as long as these frequencies are confined to the specific band being used by the phone. The receiver at the other end of the connection uses the unique identifying code that tags each packet to reassemble the message.

Because of the increased demands for data access that smartphones and similar technologies put on cell phone networks, a third generation of digital transmission technology – often referred to as “3G” – was created, which includes a variety of features to facilitate faster data transfer and to handle larger multimedia files. 3G is not in and of itself a cellular standard, but rather a group of technologies that conform to specifications issued by the International Telecommunication Union (ITU). Widely used 3G technologies include the UMTS (“Universal Mobile Telecommunications System”) system, used in Europe, Japan, and China, and EDGE, DECT, and CDMA2000, used in the United States and South Korea. Most of these technologies are backwards compatible with earlier 2G technologies, ensuring interoperability between handsets.

4G systems are the newest standards and include LTE (“Long-Term Evolution”) and WiMAX standards. Besides offering much faster data transmission rates, 4G systems can move much larger packets of data much faster; they also operate on totally different frequency bands

than previous digital cell phone systems. Each of these 4G systems builds upon modifications to previous technologies; WiMAX, for example, uses OFDM (orthogonal frequency division multiplexing), a technique similar to CDMA that divides data into multiple channels with the transmission recombined at the destination. However, WiMAX was a standard built from scratch, and has proven slow and difficult to deploy, given the expense of building new infrastructure. Many industry observers, however, see LTE as the first standard that could be adopted universally, given its flexibility and ability to operate on a wide range of radio bands (from 700 MHz to 2.6 GHz); furthermore, LTE could build upon existing infrastructure, potentially reaching a much wider range of users in short order.

Data Access and the Telecom Industry

Modern smartphones require robust, high-speed, and consistent access to the Internet in order for users to take full advantage of all of their features; as these devices have increased significantly in popularity, the rollout of advanced technologies such as 4G LTE has accelerated in recent years. Indeed, it is not uncommon for networks to market themselves on the basis of the strengths of their respective networks; commercials touting network superiority are not at all uncommon.

Despite these advances and the ongoing development of improved digital transmission technologies via radio, which has produced a significant reduction in the costs associated with operating cellular networks, American cellular companies continue to charge relatively large amounts for access to their networks, particularly when compared to their European counterparts. Most telecom companies charge smartphone users a data fee on top of their monthly cellular service charges simply for the “right” to access data, despite the fact that smartphones necessarily require data access in order to fully function. Consider this example: as of this writing, for a new smartphone subscriber, AT&T charges \$25 per month for access to just 1 gigabyte of data (extra fees are charged if one exceeds this allotment) on top of

monthly subscription fees. Prices rapidly escalate from there; 6 gigabytes costs \$80 a month, and 20 GB go for \$150 a month. Lower prices might be had if a customer is able to find a service that offers a “pay as you go” model rather than a contractual agreement; however, these types of services have some downsides, may not be as much of a bargain as advertised, and may not fully take advantage of smartphone capabilities. Technology journalist Rick Broida, for example, notes that MMS (multimedia service) messaging and visual voicemail do not work on certain no-contract carriers.

Many customers outside of the United States, on the other hand, particularly those in Europe, purchase their handsets individually and can freely choose which carrier’s SIM card to install; as a result, data prices are much lower and extremely competitive (though handsets themselves are much more expensive, as American carriers are able to largely subsidize handset costs by committing users to multi-year contracts). In fact, the European Union addressed data roaming charges in recent years and has put caps in place; as a result, as of July 1, 2014, companies may not charge more than €0.20 (approximately US\$0.26) per megabyte for cellular data. Companies are free to offer lower rates, and many do; travel writer Donald Strachan notes that a number of prepaid SIM cards can be had that offer 2 gigabytes of data for as little as €10 (approximately US\$13).

Data from Data

The digitalization of cell phones has had another consequence: digital cell phones interacting with digital networks produce a tremendous amount of data that can be analyzed for a variety of purposes. In particular, call detail records, or CDRs, are generated automatically when a cell phone connects to a network and receives or transmits; despite what the name might suggest, CDRs are generated for both text messages as well as phone calls. CDRs contain a variety of metadata related to a call including the phone numbers of the originator and receiver, call time and duration,

routing information, and so forth and are often audited by wireless networks to facilitate billing and to identify weaknesses in infrastructure.

Law enforcement agencies have long used CDRs to identify suspects, corroborate alibis, reveal behavior patterns, and establish associations with other individuals, but more recently, scholars have begun to use CDRs as a data source that can reveal information about populations. Becker et al., for example, used anonymized CDRs from cell phone users in Los Angeles, New York, and San Francisco to better understand a variety of behaviors related to human mobility, including daily travel habits, traffic patterns, and carbon emission generation; indeed, this type of work could have significant implications for urban planning, mass transit planning, alleviation of traffic congestion, combatting of carbon emissions, and more.

That said, CDRs are not the only digital “fingerprints” that cell phone users – and particularly smartphone users – leave behind as they use apps, messaging services, and the World Wide Web via their phones. Users, even those that are not knowingly or actively generating content, nevertheless create enormous amounts of information in the form of Instagram posts, Twitter messages, emails, Facebook posts, and more, virtually all of which can be identified as having originated from a smartphone thanks to meta tags and other hidden data markers (e.g., EXIF data). In some cases, these hidden data are quite extensive and can include such information as the user’s geolocation at the date and time of posting, the specific platform used, and more. Furthermore, much of these data can be harvested and examined without individual users ever knowing; Twitter sentiment analysis, for example, can be conducted specifically on messages generated via mobile platforms.

Passive collection of data generated by cell phones is not the only method available for studying cell phone data. Some intrepid researchers and research firms, recognizing that cell phones are a rich source of information on people, their interpersonal interactions, and their mobility, have developed various pieces of software that can (voluntarily) be installed on

smartphones to facilitate the tracking of subjects and their behaviors. Such studies have included the collection of communiqués as well as proximity data; in fact, researchers have found that this type of data alone is, in many cases, enough to infer friendships and close interpersonal relationships between individuals (see Eagle, Pentland, and Lazer for one example). The possibilities for such research driven by data collected via user-installed apps are potentially limitless, for both academia and industry; however, as such efforts undoubtedly increase, it will be important to ensure that end users are fully aware of the risks inherent in disclosing personal information and that users have fully consented to participation in data collection activities.

Cross-References

- [Cell Phone Data](#)
- [Data Mining](#)
- [Network Data](#)

Further Reading

- Ahmad, A. (2005). *Wireless and mobile data networks*. Hoboken: Wiley.
- Becker, R., et al. (2013). Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1), 74. <https://doi.org/10.1145/2398356.2398375>.
- Broida, R. Should you switch to a no-contract phone carrier? *CNET*. <http://www.cnet.com/news/have-you-tried-a-no-contract-phone-carrier/>. Accessed Sept 2014.
- Cox, C. (2012). *An introduction to LTE: LTE, LTE-advanced, SAE and 4G mobile communications*. Chichester: Wiley.
- Eagle, N., Pentland, A. (Sandy), & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36), 15274. <https://doi.org/10.1073/pnas.0900282106>.
- Gibson, J. D. (Ed.). (2013). *Mobile communications handbook* (3rd ed.). Boca Raton: CRC Press.
- Mishra, A. R. (2010). *Cellular technologies for emerging markets: 2G, 3G and beyond*. Chichester: Wiley.
- Strachan, D. The best local SIM cards in Europe. *The Telegraph*. <http://www.telegraph.co.uk/travel/travel-advice/9432416/The-best-local-SIM-cards-in-Europe.html>. Accessed Sept 2014.
- Yi, S. J., Chun, S. D., Lee, Y. D., Park, S. J., & Jung, S. H. (2012). *Radio protocols for LTE and LTE-advanced*. Singapore: Wiley.
- Zhang, Y., & Årvidsson, A. (2012). Understanding the characteristics of cellular data traffic. *ACM SIGCOMM Computer Communication Review*, 42(4), 461. <https://doi.org/10.1145/2377677.2377764>.

Census Bureau (U.S.)

Stephen D. Simon
P. Mean Consulting, Leawood, KS, USA

The United States Bureau of the Census (hereafter Census Bureau) is a federal agency that produces big data that is of direct value and which also provides the foundation for analyses of other big data sources. They also produce information critical for geographic information systems in the United States.

The Census Bureau is mandated by Article I, Section II of the US Constitution to enumerate the population of the United States to allow the proper allocation of members of the House of Representatives to each state. This census was first held in 1790 and then every 10 years afterwards. Full data from the census is released 72 years after the census was held. With careful linking across multiple censuses, researchers can track individuals such as Civil War veterans (Costa et al. 2017) across their full lifespan, or measure demographic changes in narrowly defined geographic regions, such as marriage rates during the boll weevil epidemic of the early 1900s (Bloome et al. 2017).

For more recent censuses, samples of microdata are available, though with steps taken to protect confidentiality (Dreschler and Reiter 2012). Information from these sources as well as census microdata from 79 other countries is available in a standardized format through the Integrated Public Use Microdata Series International Partnership (Ruggles et al. 2015).

Starting in 1940, the Census Bureau asked additional questions for a subsample of the census. The questions, known informally as “the long form,” had questions about income, occupation,

education, and other socioeconomic issues. In 2006, the long form was replaced with the American Community Survey (ACS), which covered similar issues, but which was run continuously rather than once every 10 years (Torrieri 2007). The ACS has advantages associated with the timeliness of the data, but some precision was lost compared to the long form (Spielman et al. 2014; Macdonald 2006).

Both the decennial census and the ACS rely on the Master Address File (MAF), a list of all the addresses in the United States where people might live. The MAF is maintained and updated by the Census Bureau from a variety of sources but predominantly the delivery sequence file of the United States Postal Service (Loudermilk and Li 2009).

Data from the MAF are aggregated into contiguous geographic regions. The regions are chosen to follow, whenever possible, permanent visible features like streets, rivers, and railroads and to avoid crossing county or state lines, with the exception of regions within Indian reservations (Torrieri 1994, Chapter 10). The geographic regions defined by the Census Bureau have many advantages over other regions, such as those defined by zip codes (Krieger et al. 2002). Shapefiles for various census regions are available for free download from the Census Bureau website.

The census block, the smallest of these regions, typically represent what would normally be considered a city block in an urban setting, though the size might be larger for suburban and rural settings. There are many census blocks with zero reported population, largely because the areas are uninhabitable or because residence is prohibited (Freeman 2014).

Census blocks are aggregated into block groups that contain roughly 600 to 3000 people. The census block group is the smallest geographic region for which the Census Bureau provides aggregate statistics and sample microdata (Croner et al. 1996).

Census block groups are aggregated into census tracts. Census tracts are relatively homogeneous in demographics and self-contained within county boundaries or American Indian

reservations. Tracts are relatively stable over time, with merges and partitions as needed to keep the number of people in a census tract reasonably close to 4000 (Torrieri 1994, Chapter 10).

The Census Bureau also aggregates geographic regions into Metropolitan Statistical Areas (MSA), categorizes regions on an urban/rural continuum, and clusters states into broad national regions. All of these geographic regions provide a framework for many big data analyses and helps make research more uniform and replicable.

The geographic aggregation is possible because of another product that is of great value to big data applications, the Topologically Integrated Geographic Encoding and Referencing (TIGER) System (Marx 1990). The TIGER System, a database of land features like roads and rivers and administrative boundaries like county and state lines, has formed the foundation of many commercial mapping products used in big data analysis (Croner et al. 1996). The TIGER system allows many useful characterizations of geographic regions, such as whether a region contains a highway ramp, a marker of poor neighborhood quality (Freisthler et al. 2016), and whether a daycare center is near a busy road (Houston et al. 2006).

The ACS is the flagship survey of the Census Bureau and has value in and of itself, but also is important in supplementing other big data sources. The ACS is a self-report mail survey with a telephone follow up for incomplete or missing surveys. It targets roughly 300,000 households per month. Response to the ACS is mandated by law, but the Census Bureau does not enforce this mandate. The ACS releases 1 year summaries for large census regions, 3 year summaries for smaller census regions, and 5 year summaries for every census region down to the block group level. This release schedule represents the inevitable trade-off between the desire for a large sample size and the desire for up-to-date information. The ACS has been used to describe health insurance coverage (Davern et al. 2009), patterns of residential segregation (Louf and Barthelemy 2016), and disability rates (Siordia 2015). It has also been used to

supplement other big data analysis by developing neighborhood socioeconomic status covariates (Kline et al. 2017) and obtaining the denominators needed for local prevalence estimates (Grey et al. 2016). The National Academies Press has a detailed guide on how to use the ACS (Citro and Kalton 2007) available in book form or as a free PDF download.

The Census Bureau conducts many additional surveys in connection with other federal agencies. The American Housing Survey (AHS) is a joint effort with the Department of Housing and Urban Development that surveys both occupied and vacant housing units in a nationally representative sample and a separate survey of large MSAs. The AHS conducts computer-assisted interviews of roughly 47,000 housing units biennially. The AHS allows researchers to see whether mixed use development influences commuting choices (Cervero 1996) and to assess measures of the house itself (such as peeling paint) and the neighborhood (such as nearby abandoned buildings) that can correlated with health outcomes (Jacobs et al. 2009).

The Current Population Survey, a joint effort with the Bureau of Labor Statistics, is a monthly survey of 60,000 people that provides unemployment rates for the United States as a whole and for local regions and specific demographic groups. The survey includes supplements that allows for the analysis of tobacco use (Zhu et al. 2017), poverty (Pac et al. 2017), food security (Jernigan et al. 2017), and health insurance coverage (Pascale et al. 2016).

The Consumer Expenditure Survey, also a joint effort with the Bureau of Labor Statistics, is an interview survey of major expenditure components combined with a diary study of detailed individual purchases that is integrated to provide a record of all expenditures of a family. The purchasing patterns form the basis for the market basket of goods used in computation of a measure of inflation, the Consumer Price Index. Individual level data from this survey allows for detailed analysis of purchasing habits, such as expenditures in tobacco consuming households (Rogers et al. 2017) and food expenditures of different ethnic groups (Ryabov 2016).

The National Crime Victimization Survey, a joint effort with the Bureau of Justice Statistics, is a self-report survey of 160,000 households per year on nonfatal personal crimes and household property crimes. The survey has supplements for school violence (Musu-Gillette et al. 2017) and stalking (Menard and Cox 2016).

While the Census Bureau conducts its own big data analyses, it also provides a wealth of information to anyone interested in conducting large-scale nationally representative analyses. Statistics within the geographic regions defined by the Census Bureau serve as the underpinnings of analyses of many other big data sources. Finally, the Census Bureau provides free geographic information system resources through their TIGER files.

Further Reading

- Bloome, D., Feigenbaum, J., & Muller, C. (2017). Tenancy, marriage, and the boll weevil infestation, 1892–1930. *Demography*, 54(3), 1029–1049.
- Cervero, R. (1996). Mixed land-uses and commuting: Evidence from the American Housing Survey. *Transportation Research Part A: Policy and Practice*, 30(5), 361–377.
- Citro, C. F., & Kalton, G. (Eds.). (2007). *Using the American Community Survey: Benefits and challenges*. Washington, DC: The National Academies Press.
- Costa, D. L., DeSomer, H., Hanss, E., Roudiez, C., Wilson, S. E., & Yetter, N. (2017). Union army veterans, all grown up. *Historical Methods*, 50, 79–95.
- Croner, C. M., Sperling, J., & Broome, F. R. (1996). Geographic Information Systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine*, 15, 1961–1977.
- Davern, M., Quinn, B. C., Kenney, G. M., & Blewett, L. A. (2009). The American Community Survey and health insurance coverage estimates: Possibilities and challenges for health policy researchers. *Health Services Research*, 44(2 Pt 1), 593–605.
- Dreschler, J., & Reiter, J. P. (2012). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of American Statistical Association*, 105(492), 1347–1357.
- Freeman, N. M. (2014). Nobody lives here: The nearly 5 million census blocks with zero population. <http://tumblr.mapsbynik.com/post/82791188950/nobody-lives-here-the-nearly-5-million-census>. Accessed 6 Aug 2017.
- Freisthler, B., Ponicki, W. R., Gaidus, A., & Gruenewald, P. J. (2016). A micro-temporal geospatial analysis of medical marijuana dispensaries and crime in Long Beach California. *Addiction*, 111(6), 1027–1035.

- Grey, J. A., Bernstein, K. T., Sullivan, P. S., Purcell, D. W., Chesson, H. W., Gift, T. L., & Rosenberg, E. S. (2016). Estimating the population sizes of men who have sex with men in US states and counties using data from the American Community Survey. *JMIR Public Health Surveill*, 2(1), e14.
- Houston, D., Ong, P. M., Wu, J., & Winer, A. (2006). Proximity of licensed childcare to near-roadway vehicle pollution. *American Journal of Public Health*, 96(9), 1611–1617.
- Jacobs, D., Wilson, J., Dixon, S. L., Smith, J., & Evens, A. (2009). The relationship of housing and population health: A 30-year retrospective analysis. *Environmental Health Perspectives*, 117(4), 597–604.
- Jernigan, V. B. B., Huyser, K. R., Valdes, J., & Simonds, V. W. (2017). Food insecurity among American Indians and Alaska Natives: A national profile using the current population survey-food security supplement. *Journal of Hunger and Environmental Nutrition*, 12(1), 1–10.
- Kline, K., Hadler, J. L., Yousey-Hindes, K., Niccolai, L., Kirley, P. D., Miller, L., Anderson, E. J., Monroe, M. L., Bohm, S. R., Lynfield, R., Bargsten, M., Zansky, S. M., Lung, K., Thomas, A. R., Brady, D., Schaffner, W., Reed, G., & Garg, S. (2017). Impact of pregnancy on observed sex disparities among adults hospitalized with laboratory-confirmed influenza, FluSurv-NET, 2010–2012. *Influenza and Other Respiratory Viruses*, 11(5), 404–411.
- Krieger, N., Waterman, P., Chen, J. T., Soobader, M. J., Subramanian, S. V., & Carson, R. (2002). Zip code caveat: Bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas – The Public Health Disparities Geocoding Project. *American Journal of Public Health*, 92(7), 1100–1102.
- Loudermilk, C. L., & Li, M. (2009). A national evaluation of coverage for a sampling frame based on the Master Address File. *Proceedings of the Joint Statistical Meeting*. American Statistical Association, Alexandria, VA.
- Louf, R., & Barthelemy, M. (2016). Patterns of residential segregation. *PLoS One*, 11(6), e0157476.
- Macdonald, H. (2006). The American Community Survey: Warmer (more current), but fuzzier (less precise) than the decennial census. *Journal of the American Planning Association*, 72(4), 491–503.
- Marx, R. W. (1990). The Census Bureau's TIGER system. *New Zealand Cartography Geographic Information Systems*, 17(1), 17–113.
- Menard, K. S., & Cox, A. K. (2016). Stalking victimization, labeling, and reporting: Findings from the NCVS stalking victimization supplement. *Violence Against Women*, 22(6), 671–691.
- Musu-Gillette, L., Zhang, A., Wang, K., Zhang, J., & Oudekerk, B. A. (2017). Indicators of school crime and safety: 2016. <https://www.bjs.gov/content/pub/pdf/iscs16.pdf>. Accessed 6 Aug 2017.
- Pac, J., Waldfogel, J., & Wimer, C. (2017). Poverty among foster children: Estimates using the supplemental poverty measure. *Social Service Review*, 91(1), 8–40.
- Pascale, J., Boudreaux, M., & King, R. (2016). Understanding the new current population survey health insurance questions. *Health Services Research*, 51(1), 240–261.
- Rogers, E. S., Dave, D. M., Pozen, A., Fahs, M., & Gallo, W. T. (2017). Tobacco cessation and household spending on non-tobacco goods: Results from the US Consumer Expenditure Surveys. *Tobacco Control*; pii: tobaccocontrol-2016-053424.
- Ruggles, S., McCaa, R., Sobek, M., & Cleveland, L. (2015). The IPUMS collaboration: Integrating and disseminating the world's population microdata. *Journal of Demographic Economics*, 81(2), 203–216.
- Ryabov, I. (2016). Examining the role of residential segregation in explaining racial/ethnic gaps in spending on fruit and vegetables. *Appetite*, 98, 74–79.
- Siordia, C. (2015). Disability estimates between same- and different-sex couples: Microdata from the American Community Survey (2009–2011). *Sexuality and Disability*, 33(1), 107–121.
- Spielman, S. E., Folch, D., & Nagle, N. (2014). Patterns and causes of uncertainty in the American Community Survey. *Applied Geography*, 46, 147–157.
- Torrieri, N. K. (1994). Geographic areas reference manual. <https://www.census.gov/geo/reference/garm.html>. Accessed 7 Aug 2017.
- Torrieri, N. (2007). America is changing, and so is the census: The American Community Survey. *The American Statistician*, 61(1), 16–21.
- Zhu, S. H., Zhuang, Y. L., Wong, S., Cummins, S. E., & Tedeschi, G. J. (2017). E-cigarette use and associated changes in population smoking cessation: Evidence from US current population surveys. *BMJ (Clinical Research Ed.)*, 358, j3262.

Centers for Disease Control and Prevention (CDC)

Stephen D. Simon

P. Mean Consulting, Leawood, KS, USA

The Centers for Disease Control and Prevention (CDC) is a United States government agency self-described as “the nation’s health protection agency” (CDC 2017a). CDC responds to new and emerging health threats and conducts research to track chronic and acute diseases. Of greatest interest to readers of this article are the CDC efforts in surveillance using nationwide cross-sectional surveys to monitor the health of diverse populations (Frieden 2017). These

surveys are run annually, in some cases across more than five decades. The National Center for Health Statistics (NCHS), a branch of the CDC either directly conducts or supervises the collection and storage of the data from most of these surveys.

The National Health Interview Survey (NHIS) conducts in-person interviews about the health status and health care access for 35,000 households per year, with information collected about the household as a whole and for one randomly selected adult and one randomly selected child (if one is available) in that household (Parsons et al. 2014). NHIS has been used to assess health insurance coverage (Martinez and Ward 2016), the effect of physical activity on health (Carlson et al. 2015) and the utilization of cancer screening (White et al. 2017).

The National Health and Nutrition Examination Survey (NHANES) conducts in-person interviews about the diet and health of roughly 5000 participants per year combined with a physical exam for each participant (Johnson et al. 2014). Sera, plasma, and urine are collected during the physical exam. Genetic information is extracted from the sera specimens, although consent rates among various ethnic groups are uneven (Gabriel et al. 2014). NHANES has been used to identify dietary trends in patients with diabetes (Casagrande and Cowie 2017), the relationship between inadequate hydration and obesity (Chang et al. 2016), and the association of Vitamin D levels and telomere length (Beilfuss et al. 2017).

The Behavioral Risk Factor Surveillance System (BRFSS) conducts telephone surveys of chronic conditions and health risk behaviors using random digit dialing (including cell phone numbers from 2008 onward) for 400,000 participants per year. This represents the largest telephone survey in the world (Pierannunzi et al. 2013). This survey has been used to identify time trends in asthma prevalence (Bhan et al. 2015), fall injuries among the elderly (Bergen et al. 2016), and mental health disparities between male and female caregivers (Edwards et al. 2016).

There are additional surveys of other patient populations as well as surveys of hospitals (both

inpatient and emergency room visits), physician offices, and long-term care providers. The micro-data from all of these surveys are publicly available, usually in compressed ASCII format and/or comma-separated value format. CDC also provides code in SAS, SPSS, and STATA for reading some of these files.

Since these surveys span many years, researchers can examine short- and long-term trends in health. Time trend analysis, however, does require care. The surveys can change from year to year in the sampling frame, the data collected, the coding systems, and the handling of missing values.

To improve efficiency, many of the CDC databases use a complex survey approach where geographic regions are randomly selected and then patients are selected within those regions. Often minority populations are oversampled to allow sufficient sample sizes in these groups. Both the complex survey design and the oversampling require use of specialized statistical analysis approaches (Lumley 2010; Lewis 2016).

The CDC has removed any information that could be used to personally identify individual respondents, particularly geocoding. For those researchers requiring this level of information can apply for access through the NCHS Research Data Center (CDC 2017b).

The CDC maintains the National Death Index (NDI), a centralized database of death certificate data collected from each of the 50 states and the District of Columbia. The raw data is not available for public use, but researchers can apply for access that lets them submit a file of patients that they are studying to see which ones have died (CDC 2016). Many of the CDC surveys described above are linked automatically to NDI. While privacy concerns restrict direct access to the full information on the death certificate, the CDC does offer geographically and demographically aggregated data sets on deaths (and births) as well as reduced data sets on individual deaths and births with personal identifiers removed.

The CDC uses big data in its own tracking of infectious diseases. The US Influenza Hospitalization Surveillance Network (FluSurv-Net) monitors influenza hospitalizations in 267 acute care

hospitals serving over 27 million people (Chaves et al. 2015).

Real time reporting on influenza is available at the FluView website (<https://www.cdc.gov/flu/weekly/>). This site reports on a weekly basis the outpatient visits, hospitalizations, and death rates associated with influenza. It also monitors the geographic spread, the strain type, and the drug resistance rates for influenza.

FoodNet tracks laboratory confirmed food-borne illnesses in ten geographic areas with a population of 48 million people (Crim et al. 2015). The Active Bacterial Core Surveillance collects data on invasive bacterial infections in ten states representing up to 42 million people (Langley et al. 2015).

The hallmark of every CDC data collection effort is the great care taken in either tracking down every event in the regions being studied, or in collecting a nationally representative sample. These efforts insure that researchers can extrapolate results from these surveys to the United States as a whole. These data sets, most of which are available at no charge, represent a tremendous resource to big data researchers interested in health surveillance in the United States.

Further Reading

- Beilfuss, J., Camargo, C. A. Jr, & Kamycheva, E. (2017). Serum 25-Hydroxyvitamin D has a modest positive association with leukocyte telomere length in middle-aged US adults. *Journal of Nutrition*. <https://doi.org/10.3945/jn.116.244137>.
- Bergen, G., Stevens, M. R., & Burns, E. R. (2016). Falls and fall injuries among adults aged ≥ 65 years – United States, 2014. *Morbidity and Mortality Weekly Report*, 65(37), 993–998.
- Bhan, N., Kawachi, I., Glymour, M. M., & Subramanian, S. V. (2015). Time trends in racial and ethnic disparities in asthma prevalence in the United States from the Behavioral Risk Factor Surveillance System (BRFSS) Study (1999–2011). *American Journal of Public Health*, 105(6), 1269–1275. <https://doi.org/10.2105/AJPH.2014.302172>.
- Carlson, S. A., Fulton, J. E., Pratt, M., Yang, Z., & Adams, E. K. (2015). Inadequate physical activity and health care expenditures in the United States. *Progress in Cardiovascular Disease*, 57(4), 315–323. <https://doi.org/10.1016/j.pcad.2014.08.002>.
- Casagrande, S. S., & Cowie, C. C. (2017). Trends in dietary intake among adults with type 2 diabetes: NHANES 1988–2012. *Journal of Human Nutrition and Dietetics*. <https://doi.org/10.1111/jhn.12443>.
- Centers for Disease Control and Prevention. (2016). About NCHS – NCHS fact sheets – National death index. https://www.cdc.gov/nchs/data/factsheets/factsheet_ndi.htm. Accessed 10 Mar 2017.
- Centers for Disease Control and Prevention. (2017a). Mission, role, and pledge. <https://www.cdc.gov/about/organization/mission.htm>. Accessed 13 Feb 2017.
- Centers for Disease Control and Prevention. (2017b). RDC – NCHS research data center. <https://www.cdc.gov/rdc/index.htm>. Accessed 6 Mar 2017.
- Chang, T., Ravi, N., Plegue, M. A., Sonnevile, K. R., & Davis, M. M. (2016). Inadequate hydration, BMI, and obesity among US adults: NHANES 2009–2012. *Annals of Family Medicine*, 14(4), 320–324. <https://doi.org/10.1370/afm.195>.
- Chaves, S. S., Lynfield, R., Lindegren, M. L., Bresee, J., & Finelli, L. (2015). The US influenza hospitalization surveillance network. *Emerging Infectious Diseases*, 21(9), 1543–1550. <https://doi.org/10.3201/eid2109.141912>.
- Crim, S. M., Griffin, P. M., Tauxe, R., Marder, E. P., Gilliss, D., Cronquist, A. B., et al. (2015). Preliminary incidence and trends of infection with pathogens transmitted commonly through food – Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2006–2014. *Morbidity and Mortality Weekly Report*, 64(18), 495–499.
- Edwards, V. J., Anderson, L. A., Thompson, W. W., & Deokar, A. J. (2016). Mental health differences between men and women caregivers, BRFSS 2009. *Journal of Women & Aging*. <https://doi.org/10.1080/08952841.2016.1223916>.
- Frieden, T. (2017). A safer, healthier U.S.: The centers for disease control and prevention, 2009–2016. *American Journal of Preventive Medicine*, 52(3), 263–275. <https://doi.org/10.1016/j.amepre.2016.12.024>.
- Gabriel, A., Cohen, C. C., & Sun, C. (2014). Consent to specimen storage and continuing studies by race and ethnicity: A large dataset analysis using the 2011–2012 National Health and Nutrition Examination Survey. *Scientific World Journal*. <https://doi.org/10.1155/2014/120891>.
- Johnson, C. L., Dohrmann, S. M., Burt, V. L., & Mohadjer, L. K. (2014). National health and nutrition examination survey: Sample design, 2011–2014. *Vital and Health Statistics*, 2(162).
- Langley, G., Schaffner, W., Farley, M. M., Lynfield, R., Bennett, N. M., Reingold, A. L., et al. (2015). Twenty years of active bacterial core surveillance. *Emerging Infectious Diseases*, 21(9), 1520–1528. <https://doi.org/10.3201/eid2109.141333>.
- Lewis, T. H. (2016). *Complex survey data analysis with SAS*. New York: Chapman and Hall.
- Lumley, T. (2010). *Complex surveys. A guide to analysis using R*. New York: Wiley.

- Martinez, M. E., & Ward, B. W. (2016). Health care access and utilization among adults aged 18–64, by poverty level: United States, 2013–2015. *NCHS Data Brief*, 262, 1–8.
- Parsons, V. L., Moriarity, C., Jonas, K., Moore, T. F., Davis, K. E., & Tompkins, L. (2014). Design and estimation for the national health interview survey, 2006–2015. *Vital and Health Statistics*, 165, 1–53.
- Pierannunzi, C., Hu, S. S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Medical Research Methodology*. <https://doi.org/10.1186/1471-2288-13-49>.
- White, A., Thompson, T. D., White, M. C., Sabatino, S. A., de Moor, J., Doria-Rose, P. V., et al. (2017). Cancer screening test use – United States, 2015. *Morbidity and Mortality Weekly Report*, 66(8), 201–206.

The Charter rights concern six macro areas: dignity, freedoms, equality, solidarity, citizens' rights, and justice. These six areas represent those “fundamental rights and freedoms recognized by the European Convention on Human Rights, the constitutional traditions of the EU member states, the Council of Europe’s Social Charter, the Community Charter of Fundamental Social Rights of Workers and other international conventions to which the European Union or its member states are parties” (European Parliament 2001, February 21). Europeans can use judicial and political mechanisms to hold EU institutions, and in certain circumstances, member states, accountable in those situations where they do not comply with the Charter. The Charter can be used as a political strategy to influence decision-makers so to develop policies and legislation that are in line with human right standards (Butler 2013).

Charter of Fundamental Rights (EU)

Chiara Valentini

Department of Management, Aarhus University,
School of Business and Social Sciences, Aarhus,
Denmark

Introduction

The Charter of Fundamental Rights is a legal document that protects individuals and legal entities from actions that disregard fundamental rights. It covers personal, civic, political, economic, and social rights of people within the European Union. The Charter also safeguards so-called “third generation” fundamental rights, such as data protection, bioethics, and transparent administration matters, which includes protection from the misuse of massive datasets on individuals’ online behaviors collected by organizations. Diverse organizations have taken advantage of large data sets and big data analytics to bolster competitiveness, innovation, market predictions, political campaigns, targeted advertising, scientific research, and policymaking and to influence elections and political outcomes through, for instance, targeted communications (European Parliament 2017, February 20).

Historical Development

The Charter was drafted during the European Convention and was solemnly proclaimed by the three major EU decision-making institutions, that is, the European Parliament, the Council of the European Union and the European Commission, on December 7, 2000. Before the Charter was written, the EU had already internal rules on human rights, but these were not incorporated in a legal document. They were only part of the general principles governing EU law. In practice, the lack of a legal document systematically addressing questions of human rights permitted a number of EU law infringements. For instance, in situations where member states needed to transpose EU law into their national ones, in some cases national courts refused to apply EU law with the content that it conflicted with rights protected by their national constitutions (Butler 2013). To solve the issue of EU law infringement as well as to harmonize EU legislations in relation to fundamental rights, the European Council entrusted a group of people during the 1999 Cologne meeting to form the European Convention, a body set up to deal with the task of

drafting the Charter of Fundamental Rights (Nugget 2010; Butler 2013).

The endorsement of respecting the Charter of Fundamental Rights by the three major EU political institutions sparked a new political discussion on whether the Charter should be included in the EU Constitutional Treaty, which was at the top of the political agenda in early 2000, and on whether the EU should sign up to the European Convention of Human Rights (ECHR). The Charter was amended a number of times and ended up not to be included in the Constitutional Treaty (Nugget 2010). Nonetheless several of the underpinning rights became legally binding with the entry into force of the Lisbon Treaty in December 2009. De facto, the Charter has gained some legal impact in EU legal system. Today the Charter can regulate the activities of national authorities that implement EU laws at national level. But it cannot be used in cases of infringements of rights for actions dealing with national legislation. The Charter has also limited influence in certain countries that have obtained some opt-outs. Member states that are granted opt-outs are allowed not to implement certain EU policies. In March 2017, the European Parliament voted on a nonlegislative resolution about the fundamental rights implications of big data, including privacy, data protection, nondiscrimination, security, and law-enforcement. Essentially, the resolution tries to address recommendations for digital literacy, ethical frameworks, and guidelines for algorithmic accountability and transparency. It also seeks to foster cooperation among authorities, regulators, and the private sector and to promote the use of security measures like privacy by design and by default, anonymization techniques, encryption, and mandatory privacy impact assessments (European Parliament 2017, February 20).

Surveillance Practices and Protection of Human Rights

Because the rights in the Charter are binding EU legislation, the European Parliament, the Council

of the European Union and the European Commissions have specialized bodies and procedures to help ensuring that proposals are consistent with the Charter (Butler 2013). Furthermore, to increase awareness and knowledge on fundamental rights, the European Council decided on 13 December 2003 to extend the duties of an existing agency, the European Monitoring Centre on Racism and Xenophobia, to include the monitoring of human rights. A newly formed community agency, the European Union Agency for Fundamental Rights (FRA), was established in 2007 and is based in Vienna, Austria (CEC 2007, February 22). As declared, the main scope of this agency is “to collect and disseminate objective, reliable and comparable data on the situation of fundamental rights in all EU countries within the scope of EU law” (European Commission 2013, July 16).

During the past years, the agency has been involved in investigating the status of surveillance practices in Europe, specifically in relation to the respect for private and family rights and the protection of personal data (FRA 2014, May). The agency receives the mandate to carry out investigations by the European Parliament. The scope is gathering information on the status of privacy, security, and transparency in the EU. Specifically, the agency appraises when, how, and for which purposes member states collect data on the content of communications and metadata and follow citizens’ electronic activities, in particular in their use of smartphones, tablets, and computers (European Parliament 2014, February 21). In 2014, the agency found out that mass surveillance programs were in place in some member states breaching EU fundamental rights (LIBE 2014). On the basis of the agency’s investigative work, the European Parliament voted a resolution addressing the issue of mass surveillance. The agency continues recognizing the importance of big data for today’s information society as a way for boosting innovation, yet it acknowledges the importance of finding a right balance between the challenges linked to security and respect of fundamental rights by helping EU policymakers and its members states with

updated research on how large sets of data collections are conducted within the EU.

Cross-References

- European Commission
- European Commission: Directorate-General for Justice (Data Protection Division)
- European Union

Further Reading

- Butler, I. (2013). The European charter of fundamental right: What can I do?. *Background paper of the open society European Policy Institute*. <http://www.opensocietyfoundations.org/sites/default/files/eu-charter-fundamental-rights-20130221.pdf>. Accessed on 10 Oct 2014.
- CEC. (2007, February 22). Council Regulation (EC) No. 168/2007 of 15 February 2007 establishing a European Union Agency for Fundamental Rights. http://fra.europa.eu/sites/default/files/fra_uploads/74-reg_168-2007_en.pdf. Accessed on 10 Oct 2014.
- European Commission. (2013, July. 16). The European Union agency for fundamental rights. http://ec.europa.eu/justice/fundamental-rights/agency/index_en.htm. Accessed on 10 Oct 2014.
- European Parliament. (2001, February 21). The charter of fundamental rights of the European Union. http://www.europarl.europa.eu/charter/default_en.htm. Accessed on 10 Oct 2014.
- European Parliament. (2014, February 21). Report on the US NSA surveillance programme, surveillance bodies in various Member States and their impact on EU citizens' fundamental rights and on transatlantic cooperation in Justice and Home Affairs. <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A7-2014-0139+0+DOC+PDF+V0/EN>. Accessed on 10 Oct 2014.
- European Parliament. (2017, February 20). Report on fundamental rights implications of big data: Privacy, data protection, non-discrimination, security and law-enforcement (2016/2225- INI). <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0044+0+DOC+PDF+V0/EN>. Accessed on 20 June 2017.
- FRA. (2014, May). National intelligence authorities and surveillance in the EU: Fundamental rights safeguards and remedies. <http://fra.europa.eu/en/project/2014/national-intelligence-authorities-and-surveillance-eu-fundamental-rights-safeguards-and>. Accessed on 10 Oct 2014.
- LIBE. (2014). Libe Committee inquiry. Electronic mass surveillance of EU citizens. Protecting fundamental rights in a digital age. Proceedings, outcome and background documents. Document of the European Parliament, <http://www.europarl.europa.eu/document/activities/cont/201410/20141016ATT91322/20141016ATT91322EN.pdf>. Accessed on 31 Oct 2014.
- Nugget, N. (2010). *The government and politics of the European Union* (7th ed.). New York: Palgrave Macmillan.

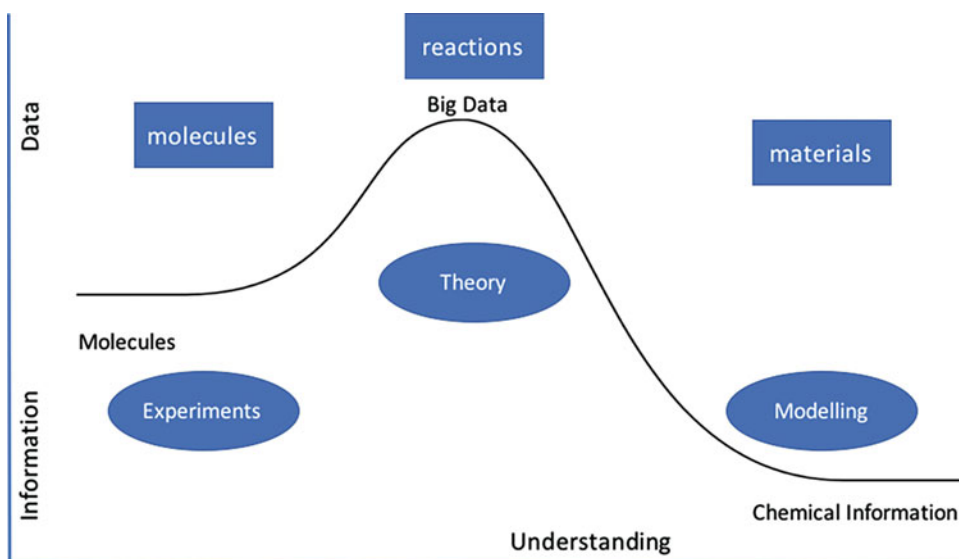
Chemistry

Colin L. Bird and Jeremy G. Frey
Department of Chemistry, University of
Southampton, Southampton, UK

Chemistry has always been data-dependent, but as computing power has increased, chemical science has become increasingly data-intensive, a development recognized by several contributors to the book edited by Hey, Tansley, and Tolle, *The Fourth Paradigm* (Hey et al. 2009). In one article, chemistry is given as one example of “a genuinely new kind of computationally driven, interconnected, Web-enabled science.”

The study of chemistry can be perceived as endeavoring to obtain big information – big in the sense of significance – from data relating to molecules, which are small in the physical sense. The transition from data to big information is perhaps well illustrated by the role of statistical mechanics as we see the move from modeling departures from ideal gas behavior through to the measurement of single molecule properties: a journey of simple information about lots of similar molecules to complex information about individual molecules, paralleled in the development of machine learning from large data sets (Ramakrishnan et al. 2015; Barrett and Langdon 2006) (Fig. 1).

As the amounts of data available have increased to the extent that chemists now handle Big Data routinely, the influence of that data tracks the evolution of the topics or disciplines



Chemistry, Fig. 1 Aspects of **Big Data Chemistry** – highlighting the rise of Big Data, the velocity (reactivity) of Big Data and the ultimate extraction of a small amount of knowledge from this data

of chemometrics and cheminformatics. Starting from the application of statistical methods to the analysis and modeling of small but critical data, chemical science has moved with the increasing quantity and complexity of chemical data, to the creation of a chemical informatics discipline to handle the need to link complex heterogeneous data.

Chemists were generating large quantities of data before the end of the twentieth century, for example, with combinatorial chemistry experiments (Lowe 1995) and, at the beginning of this century, e-Science techniques opened up prospects of greater diversity in what was achievable. High throughput chemistry created yet more data at greater speed. The term Big Data has been employed for about the same length of time, although only more recently has it been used in chemistry. Combinatorial and high throughput methods led the way for chemists to work with even greater data volumes, the challenge being to make effective use of the flood of data, although there is a point of view that chemical data, although diverse and heterogeneous, is not necessarily “big” data.

In terms of the Gartner definition (Gartner), chemical data is high-variety, can be high-

volume, and sometimes high-velocity. Four other “Vs” are also relevant to chemistry data: value, veracity, visualization, and virtual. The challenge for chemists is to make effective use of the broad range of information available from multiple sources, taking account of all seven “Vs”. Chemists therefore take a “big picture” view of their data, whatever its numerical size. Pence and Williams point out the distinction between big databases and Big Data, noting that students in particular are familiar with searching the former but may be unaware of the latter and its significance (Pence and Williams 2016).

One potentially simplifying aspect of many large sets of complex and linked chemical data relating to physical and social processes is their characterization by power-law distributions. This has not been very apparent in chemical data so far, perhaps because of limited data volumes or bias within the collection of the data, however, some interesting distributions have been suggested and this may lead to new approaches to curation and searching of large sets of linked chemical data (Benz et al. 2008).

The concept of chemical space puts into perspective not only the gamut of chemical structures but also the scale and scope of the associated

chemical data. When chemists define a space with a set of descriptors, that space will be multi-dimensional and will comprise a large number of molecules of interest, resulting in a big dataset. For example, a chemical space often referred to in medicinal chemistry is that of potential pharmacologically active molecules. The size, even when limited to small drug-like molecules, is huge, estimates vary from 10^{30} – 10^{60} (the GDB-17 (up to 17 atoms of C, N, O, S, and halogens.) enumeration of chemical space contains 166.4 billion molecules (Reymond 2015)).

However, a key question for Big Data analytics is for how many of these molecules do we have any useable data? For comparison, the Cambridge Structural Database (CSD) has 875,000 curated structures (<https://www.ccdc.cam.ac.uk>), and the Protein Data Bank (PDB) over 130,000 structures (<https://www.rcsb.org/pdb/statistics/holdings.do>), PubChem (Bolton et al. 2016) reaches into the Big Data regime as it contains over 93 million compounds but with differing quantity and quality of information for these entries (<https://pubchem.ncbi.nlm.nih.gov/>). Other comparisons of chemical data repositories are given in Tetko et al. (2016) and Bolstad et al. (2012).

The strategies for navigating the chemical space of potential molecules for new and different molecular systems has led to the development of a number of search/navigation strategies to cope with both the graphical and virtual nature (Sayle et al. 2013; Hall et al. 2017) and size of the databases (Andersen et al. 2014).

Much of the discussion about the size of Chemical space focuses on the molecular structures, but an even greater complexity is apparent in the discussion of the reactions that link these molecules and many chemists desire the data to be bigger than it currently is. This is an area where the extraction of information from the literature is crucial (Schneider et al. 2016; Jessop et al. 2011; Swain and Cole 2016). The limited publication of reactions that do not work well, or the reporting of only a selection of reaction conditions investigated, has been and will continue to be a major hindrance in the modeling of chemical reactivity [see for example the Dial-a-Molecule data initiative (<http://generic.wordpress.soton.ac.uk/dial-a-molecule/phase-iii-themes/data-driven-synthesis/>)].

Chemists, irrespective of the domain in which they are interested, will be concerned with many aspects: the preservation and curation of chemical data sets; accessing and searching for chemical data; analysis of the data, including interpreting spectroscopic data; and also with more specific information, such as thermodynamic data.

Obtaining value from high volumes of varied and heterogeneous data, sometimes at high velocity, requires “cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation (Gartner).” Chemical data might be multi-dimensional, might be virtual, might be of dubious validity, and might be difficult to visualize. The seven “Vs” appear again.

One approach that has gained popularity in recent years is to use Semantic Web techniques to deal programmatically with the meaning of the data, whether “Big” or of a more easily managed size (Frey and Bird 2013). However, chemists must acknowledge that such techniques, while potentially of great value, do lead to a major increase in the scale and complexity of the data held.

Medicinal chemistry, which comprises drug discovery, characterization, and toxicology, is the high-value field most commonly associated with Big Data in chemistry. The chemical space of candidate compounds contains billions of molecules, the efficient scanning of which requires advanced techniques, exploiting the broad variety of data sources. Data quality (veracity) is important. This vast chemical space is represented by large databases consisting of virtual compounds that require intelligent search strategies and smart visualization techniques.

The future of drug studies is likely to lie in the integration of chemical, biological (molecular biology, genomics, metabolomics) (Bon and Waldmann 2010; Araki et al. 2008; Butte and Chen 2016; Tormay 2015; Dekker et al. 2013), materials, and environmental datasets (Buytaert et al. 2015), with toxicological predictions of new materials being a very significant application of Big Data (Hartung 2016). The validation of this

data (experimental or calculated) is an important problem which is exacerbated by the data volumes but potentially ameliorated by the potential comparisons between datasets that are opened up with increasing amounts of linked data. Automated structure validation (Spek 2009), automated application of thermodynamics aided by ThermoML (<https://www.nist.gov/mml/acmd/trc/thermoml>) and potentially a change in approach to publicly available large datasets earlier (Ekins et al. 2012; Gilson et al. 2012) and databases cited earlier will be influential.

Drug design and discovery is now highly dependent on web-based services, as reviewed by Frey and Bird in 2011 (Frey and Bird 2011). The article highlighted the growing need for services and systems able to cope with the vast amounts of chemical, biochemical, and bioinformatics information.

Tools for the manipulation of chemical data are very important. Open Babel (Banck et al. 2011) is one of the leading tools which converts over 110 different formats and the open source tools CDK (Han et al. 2003) and RDKit (<http://www.RDKit.org>) are widely used in the exploration of chemical data, but further work on tools optimized for very large datasets is ongoing.

Environmental chemistry is another field in which large volumes of data are processed as we tackle the issues raised by climate change and seek ways to mitigate the effect. Edwards et al. define big environmental data as large or complex sets of structured or unstructured data that might relate to an environmental issue (Edwards et al. 2015). They note that multiple and/or varied datasets might be required, presenting challenges for analysis and visualization.

Toxicology testing is as vital for environmental compounds as for drug candidates, and similar high throughput methods are used (Zhu et al. 2014), which is providing plenty of data for subsequent model building. Efficient toxicology testing of candidate compounds is of prime importance. Automated high throughput screening techniques make feasible the in vitro testing of up to 100,000 compounds per day (Szymański et al. 2012), thereby generating large amounts of data that requires rapid – high velocity – analysis.

Mohimani et al. have recently reported a new technique that enables high-throughput scanning of mass spectra to identify potential antibiotics and other bioactive peptides (Mohimani et al. 2017).

Computational chemistry uses modeling, simulation, and complex calculations to enable chemists to understand the structure, properties, and behavior of molecules and compounds. As such it is very much dependent on data, which might be virtual, varied, and high in volume. Molecular Dynamics simulations present significant Big Data challenges (Yeguas and Casado 2014). Pence and Williams (2016) note that Big Data tools can enhance collaboration between separate research groups. The EU-funded BIGCHEM project considers how data sharing can operate effectively in the context of the pharmaceutical industry, which “mainly aims to develop computational methods specifically for Big Data analysis”, facilitating data sharing across institutions and companies (Tetko et al. 2016).

The Chemical Science community makes extensive use of large scale science research infrastructures (e.g., Synchrotron, Neutron, Laser) and the next generation of such facilities, are already coming online (e.g., LCLS (<https://lcls.slac.stanford.edu/>), XFEL (<https://www.xfel.eu/>)) as are the massive raw datasets generated by cryo-EM (Belianinov et al. 2015). These experiments are generating data on a scale that will challenge even the data produced by CERN (<https://home.cern/>). The social and computational infrastructures to deal with these new levels of production of data are the next challenge faced by the community.

The chemical industry recognizes that harnessing Big Data will be vital for every sector, the two areas of particular interest being pricing strategy and market forecasting (ICIS Chemical Business 2013). Moreover, Big Data is seen as valuable in the search for new products that cause less emission or pollution (Lundia 2015). A recent article published by consultants KPMG asserts that the “global chemical industry has reached a tipping point” (Kaestner 2016). The article suggests that data and analytics should now be considered among the pillars of a modern business.

By using analytics to integrate information from a range of sources, manufacturers can increase efficiency and improve quality: “In developed markets, companies can use Big Data to reduce costs and deliver greater innovation in products and services.”

The importance of educating present and future chemists about Big Data is gaining increased recognition. Pence and Williams argue that Big Data issues now pervade chemical research and industry to an extent that the topic should become a mandatory part of the undergraduate curriculum (2016). They acknowledge the difficulties of fitting a new topic into an already full curriculum, but believe it sufficiently important that the addition is necessary, suggesting that a chemical literature course might provide a medium.

Big Data in Chemistry should be seen in the context of the challenges and opportunities in the wider physical sciences (Clarke et al. 2016). Owing to limited space in this article, we have concentrated on the use of chemical Big Data for molecular chemistry. The complexity of the wider discussion of polymer and materials chemistry is elegantly illustrated by the discussion of the “Chemical gardens” which as the authors state, “are perhaps the best example in chemistry of a self-organizing non-equilibrium process that creates complex structures” (Barge et al. 2015), and in this light, Whitesides highlights that “Chemistry is in a period of change, from an era focused on molecules and reactions, to one in which manipulations of systems of molecules and reactions will be essential parts of controlling larger systems” (Whitesides 2015).

Further Reading

- Andersen, J. L., Flamm, C., Merkle, D., & Stadler, P. F. (2014). Generic strategies for chemical space exploration. *International Journal of Computational Biology and Drug Design*, 7(2–3), 225–258.
- Araki, M., Gutteridge, A., Honda, W., Kanehisa, M., & Yamanishi, Y. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13), i232–i240.
- Bank, M., Hutchison, G. R., James, C. A., Morley, C., O’Boyle, N. M., & Vandermeersch, T. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3, 33.
- Barge, L. M., Cardoso, S. S., Cartwright, J. H., Cooper, G. J., Cronin, L., Doloboff, I. J., Escribano, B., Goldstein, R. E., Haudin, F., Jones, D. E., Mackay, A. L., Maselko, J., Pagano, J. J., Pantaleone, J., Russell, M. J., Sainz-Diaz, C. I., Steinbock, O., Stone, D. A., Tanimoto, Y., Thomas, N. L., & Wit, A. D. (2015). From chemical gardens to chemobionics. *Chemical Reviews*, 115(16), 8652–8703.
- Barrett, S. J., & Langdon, W. B. (2006). Advances in the application of machine learning techniques in drug discovery, design and development. In A. Tiwari, R. Roy, J. Knowles, E. Avineri, & K. Dahal (Eds.), *Applications of soft computing. Advances in intelligent and soft computing* (Vol. 36). Berlin/Heidelberg: Springer.
- Belianinov, A., et al. (2015). Big data and deep data in scanning and electron microscopies: Deriving functionality from multidimensional data sets. *Advanced Structural and Chemical Imaging*, 1, 6. <https://doi.org/10.1186/s40679-015-0006-6>.
- Benz, R. W., Baldi, P., & Swamidass, S. J. (2008). Discovery of power-laws in chemical space. *Journal of Chemical Information and Modeling*, 48(6), 1138–1151.
- Bolstad, E. S., Coleman, R. G., Irwin, J. J., Mysinger, M. M., & Sterling, T. (2012). ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7), 1757–1768.
- Bolton, E., Bryant, S. H., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Kim, S., Shoemaker, B. A., Thiessen, P. A., Wang, J., Yu, B., & Zhang, J. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, 44, D1202–D1213.
- Bon, R. S., & Waldmann, H. (2010). Bioactivity-guided navigation of chemical space. *Accounts of Chemical Research*, 43(8), 1103–1114.
- Butte, A., & Chen, B. (2016). Leveraging big data to transform target selection and drug discovery. *Clinical Pharmacology and Therapeutics*, 99(3), 285–297.
- Buytaert, W., El-khatib, Y., Macleod, C. J., Reusser, D., & Vitolo, C. (2015). Web technologies for environmental Big Data. *Environmental Modelling and Software*, 63, 185–198.
- Clarke, P., Coveney, P. V., Heavens, A. F., Jäykkä, J., Kom, A., Mann, R. G., McEwen, J. D., Ridder, S. D., Roberts, S., Scanlon, T., Shellard, E. P., Yates, J. A., & Royal Society (2016). <https://doi.org/10.1098/rsta.2016.0153>.
- Dekker, A., Ennis, M., Hastings, J., Harsha, B., Kale, N., Matos, P. D., Muthukrishnan, V., Owen, G., Steinbeck, C., Turner, S., & Williams, M. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Research*, 41, D456–D463.
- Edwards, M., Aldea, M., & Belisle, M. (2015). Big Data is changing the environmental sciences. *Environmental Perspectives*, 1. Available from http://www.exponent.com/files/Uploads/Documents/Newsletters/EP_2015_Vol1.pdf.

- Ekins, S., Tkachenko, V., & Williams, A. J. (2012). Towards a gold standard: Regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discovery Today*, 17(13–14), 685–701.
- Frey, J. G., & Bird, C. L. (2011). Web-based services for drug design and discovery. *Expert Opinion on Drug Discovery*, 6(9), 885–895.
- Frey, J. G., & Bird, C. L. (2013). Cheminformatics and the semantic web: Adding value with linked data and enhanced provenance. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(5), 465–481. <https://doi.org/10.1002/wcms.1127>.
- Gartner. From the Gartner IT glossary: What is Big Data? Available from <https://www.gartner.com/it-glossary/big-data>.
- Gilson, M. K., Liu, T., & Nicola, G. (2012). Public domain databases for medicinal chemistry. *Journal of Medicinal Chemistry*, 55(16), 6987–7002.
- Groth, P. T., Gray, A. J., Goble, C. A., Harland, L., Loizou, A., & Pettifer, S. (2014). API-centric linked data integration: The open phacts discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 29, 12–18.
- Hall, R. J., Murray, C. W., & Verdonk, M. L. (2017). The fragment network: A chemistry recommendation engine built using a graph database. *Journal of Medicinal Chemistry*, 60(14), 6440–6450. <https://doi.org/10.1021/acs.jmedchem.7b00809>.
- Han, Y., Horlacher, O., Kuhn, S., Luttmann, E., Steinbeck, C., & Willighagen, E. L. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493–500.
- Hartung, T. (2016). Making big sense from big data in toxicology by read-across. *ALTEX*, 33(2), 83–93.
- Hey, A., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm, data-intensive scientific discovery*. Redmond: Microsoft Research. ISBN 978-0-9825442-0-4. <http://generic.wordpress.soton.ac.uk/dial-a-molecule/phase-iii-themes/data-driven-synthesis/>. Accessed 30 Oct 2017.
- <https://home.cern/>. Accessed 30 Oct 2017.
- <https://lcls.slac.stanford.edu/>. Accessed 30 Oct 2017.
- <https://pubchem.ncbi.nlm.nih.gov/>. Accessed 30 Oct 2017.
- <https://www.ccdc.cam.ac.uk>. Accessed 30 Oct 2017.
- <https://www.nist.gov/mml/acmd/trc/thermoml>. Accessed 30 Oct 2017.
- <http://www.RDKit.org>. Accessed 30 Oct 2017.
- <https://www.rcsb.org/pdb/statistics/holdings.do>. Accessed 30 Oct 2017.
- <https://www.xfel.eu/>. Accessed 30 Oct 2017.
- ICIS Chemical Business. (2013). Big data and the chemical industry. Available from <https://www.icis.com/resources/news/2013/12/13/9735874/big-data-and-the-chemical-industry/>.
- Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., & Murray-Rust, P. (2011). OSCAR4: A flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3, 41. <https://doi.org/10.1186/1758-2946-3-41>.
- Kaestner, M. (2016). Big Data means big opportunities for chemical companies. *KPMG REACTION*, 16–29.
- Lowe, G. (1995). Combinatorial chemistry. *Chemical Society Review*, 24, 309–317. <https://doi.org/10.1039/CS9952400309>.
- Lundia, S. R. (2015). How big data is influencing chemical manufacturing. Available from <https://www.chem.info/blog/2015/05/how-big-data-influencing-chemical-manufacturing>.
- Mohimani, H., et al. (2017). Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology*, 13, 30–37. <https://doi.org/10.1038/nchembio.2219>.
- Pence, H. E., & Williams, A. J. (2016). Big data and chemical education. *Journal of Chemical Education*, 93(3), 504–508. <https://doi.org/10.1021/acs.jchemed.5b00524>.
- Peter V. Coveney, Edward R. Dougherty, Roger R. Highfield, (2016) Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2080):20160153
- Ramakrishnan, R., Dral, P. O., Rupp, M., & Anatole von Lilienfeld, O. (2015). Big data meets quantum chemistry approximations: The Δ -machine learning approach. *Journal of Chemical Theory and Computation*, 11(5), 2087–2096. <https://doi.org/10.1021/acs.jctc.5b00099>.
- Reymond, J. (2015). The chemical space project. *Accounts of Chemical Research*, 48(3), 722–730.
- Sayle, R. A., Batista, J., & Grant, A. (2013). An efficient maximum common subgraph(MCS) searching of large chemical databases. *Journal of Cheminformatics*, 5(1), O15. <https://doi.org/10.1186/1758-2946-5-S1-O15>.
- Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A., & Landrum, G. A. (2016). Big data from pharmaceutical patents: A computational analysis of medicinal chemists' bread and butter. *Journal of Medicinal Chemistry*, 59(9), 4385–4402. <https://doi.org/10.1021/acs.jmedchem.6b00153>.
- Spek, A. L. (2009). Structure validation in chemical crystallography. *Acta Crystallographica. Section D, Biological Crystallography*.
- Swain, M. C., & Cole, J. M. (2016). ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10), 1894–1904. <https://doi.org/10.1021/acs.jcim.6b00207>.
- Szymański, P., Marcowicz, M., & Mikiciuk-Olasik, E. (2012). Adaptation of high-throughput screening in drug discovery – Toxicological screening tests. *International Journal of Molecular Sciences*, 13, 427–452. <https://doi.org/10.3390/ijms13010427>.
- Tetko, I. V., Engkvist, O., Koch, U., Reymond, J.-L., & Chen, H. (2016). BIGCHEM: Challenges and opportunities for big data analysis in chemistry. *Molecular Informatics*, 35, 615.

- Tormay, P. (2015). Big data in pharmaceutical R&D: Creating a sustainable R&D engine. *Pharmaceutical Medicine* 29(2), 87–92.
- Whitesides, G. M. (2015). Reinventing chemistry. *Angewandte Chemie*, 54(11), 3196–3209.
- Yeguas, V., & Casado, R. (2014). *Big Data issues in computational chemistry, 2014 international conference on future internet of things and cloud*. Available from <http://ieeexplore.ieee.org/abstract/document/6984225/>.
- Zhu, H., et al. (2014). Big data in chemical toxicity research: The use of high-throughput screening assays to identify potential toxicants. *Chemical Research in Toxicology*, 27(10), 1643–1651. <https://doi.org/10.1021/tx500145h>.

Clickstream Analytics

Hans C. Schmidt
Pennsylvania State University – Brandywine,
Philadelphia, PA, USA

Clickstream analytics is a form of Web usage mining that involves the use of predictive models to analyze records of individual interactions with websites.

These records, known as clickstreams, are gathered whenever a user connects to the Web and include the totality of a Web user's browsing history. A clickstream comprises a massive amount of data, making clickstream data a type of big data. Each click, keystroke, server response, or other action is logged with a time stamp and the originating IP address, as well as information about the geographic location of individual users, the referring or previously visited website, the amount of time spent on a website, the frequency of visits by users to a website, and the purchase history of users on a website. Technical details are also gathered, and clickstream records also often include information regarding the user's Web browser, operating system, screen size, and screen resolution.

Increasingly, clickstream records also include data transmitted over smartphones, game consoles, and a variety of household appliances connected via the emerging "Internet of Things." The most comprehensive clickstream records are compiled

by Internet service providers, which provide the portal through which most individual web traffic travels. Other clickstream data are logged by both benign and malicious parties through the use of JavaScript code, CGI scripts, and tracking cookies that embed signals within individual users' Web browsers. Such passive tracking methods help to differentiate between real human interaction and machine-generated Web traffic and also automatically share clickstream data across hundreds or thousands of affiliated websites.

Clickstream data are used for a variety of purposes. One of the primary ways in which clickstream data are used involves Web advertising. By analyzing a variety of metrics, such as a website's contact efficiency, which refers to the number of unique page views, and a website's conversion efficiency, which refers to the percentage of page views that lead to purchases, advertisers can decide where to place advertisements.

Similarly, clickstream logs allow the performance of individual advertisements to be monitored by providing data about how often an advertisement is shown (impression), how often an ad is clicked on (click-through rate), and how many viewers of an ad proceed to buy the advertised product (conversion rate).

Clickstream logs are also used to create individual consumer and Web user profiles. Profiles are based on personal characteristics, such as address, gender, income, and education, and online behaviors, such as purchase history, spending history, search history, and click path history. Once profiles have been generated, they are used to help directly target advertisements, construct personalized Web search results, and create customized online shopping experiences with product recommendations.

Online advertising and e-commerce are not the only uses for clickstream analytics. Clickstream data are also used to improve website design and create faster browsing experiences. By analyzing the most frequent paths through which users navigate websites, designers can redesign Web pages to create a more intuitive browsing experience. Further, by knowing the most frequently visited websites, Internet service providers and local

network administrators can increase connectivity speeds by caching popular websites and connecting users to the locally cached pages instead of directing them to the host server.

Clickstream analytics have also come to be used by law enforcement agencies and in national defense and counterterrorism initiatives. In such instances, similar tools are employed in order to identify individuals involved with criminal, or otherwise nefarious, activities in an online environment.

While clickstream analytics have become an important tool for many organizations involved with Web-based commerce, law enforcement, and national defense, the widespread use of personal data is not without controversy.

Some object to the extensive collection and analysis of clickstream data because many Web users operate under the assumption that their actions and communications are private and anonymous and are unaware that data are constantly being collected about them. Similarly, some object because much clickstream data are collected surreptitiously by websites that do not inform visitors that data are being gathered or that tracking cookies are being used.

To this end, some organizations, like the Electronic Frontier Foundation, have advocated for increased privacy protections and suggested that Internet service providers should collect less user information. Similarly, many popular Web browsers now offer do-not-track features that are designed to limit the extent to which individual clickstreams are recorded. Yet, because there are so many points at which user data can be logged, and because technology is constantly evolving, new methods for recording online behavior continue to be developed and integrated into the infrastructure of the Web.

Cross-References

- [Data Mining](#)
- [National Security Administration \(NSA\)](#)
- [Privacy](#)

Further Reading

- Croll, A., & Power, S. (2009). *Complete web monitoring: Watching your visitors, performance, communities and competitors*. Sebastopol: O'Reilly Media.
- Jackson, S. (2009). *Cult of analytics: Driving online marketing strategies using web analytics*. Oxford: Butterworth-Heinemann.
- Kaushik, A. (2007). *Web analytics: An hour a day*. Indianapolis, IN: Wiley.

Climate Change, Hurricanes/Typhoons/Cyclones

Patrick Doupe and James H. Faghmous
Arnhold Institute for Global Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Introduction

Hurricanes, typhoons, and tropical cyclones (hereafter TCs) refer to intense storms that start at sea over warm water and sometimes reach land. Since 1970, around 90 of these storms occur annually, causing large amounts of damage (Schreck et al. 2014). The expected annual global damage from these hurricanes is US\$ 21 billion, affecting around 19 million people annually with approximately 17,000 deaths (Guha-Sapir et al. 2015). Damages from hurricanes also lower incomes, with a 90th percentile event reducing per capita incomes by 7.4% after 20 years (Hsiang and Jina 2014). This reduction is comparable with a banking crisis.

Despite these high costs, our current understanding of hurricanes is limited. For example, we lack basic understanding on *cyclogenesis*, or how cyclones form and how hurricanes link to basic variables like sea surface temperatures (SSTs). For instance, there is a positive correlation between SSTs and TC frequency in the North Atlantic Ocean but no relationship in the Pacific and Indian Ocean. Overall, the Intergovernmental Panel on Climate Change's (IPCC) current conclusion is that we have low confidence that there

are long-term robust changes in tropical cyclone activity (Pachauri et al. 2014).

Much of our limited understanding can be explained through an understanding of data. We currently have good, global data since 1970 on TC events (Schreck et al. 2014). This includes data on wind speed, air temperature, etc. The current data challenge is twofold: understanding the relationships between TCs and other variables and predicting future TCs. Progress on these topics will be made by understanding and overcoming data limitations. In this article, we show this through a review of the literature attempting to predict TCs.

Characteristics of TC Data

We can characterize TC data as being noisy, having a short-time series and large spatial dimensionality, highly autocorrelated with rare events. This, in combination with limited understanding of the physical processes, means that it is easy for researchers to overfit data points.

The data in which we do have confidence is the relatively short-time series we have on TCs (Landsea 2007; Chang and Guo 2007). The most reliable data is satellite data, which provides us with information only back to 1970. Prior to the satellite era, storm reconnaissance was through coastal or ship observations. This data is not as reliable. Given a large number of storms never reach land and that ships generally try to avoid storms, a large number of storms might have gone undetected. Another source of undercounts is low coastal population density during the earlier record (Landsea 2007). Furthermore, it has been suggested that a storm per year undercount as late as the 2003–2006 period is possible due to changes in data processing methods (Landsea 2007).

Therefore, researchers face a trade-off: longer, less reliable datasets or shorter, more reliable datasets. This trade-off is important: when we control for this observation bias, global trends in cyclone frequency become insignificant (Knutson et al. 2010).

In addition to missing data, other factors confound our estimates of the relationship between

the climate and cyclones. First and in contrast to short-time series, we have large spatial dimensionality. So although there is a reasonable amount of cyclones over time, when taking over space and time, we have few events for very many observations. Second, there are large amplitude fluctuations in present-day storms (Ding and Reiter 1981). Last, there are large knowledge gaps concerning the exact influence various climate factors have on TC activity (Gray and Brody 1967; Emanuel 2008). These characteristics of TC data mean that it is easy for researchers to overfit explanatory variables to poorly understood noise or autocorrelations in the data. We now investigate how these constraints affect cyclone forecasting.

Forecasting Cyclones

We can group TC predictions into three broad categories. First *centennial projections* are simulations used to model TC activity under various warming scenarios. Centennial projections generally look at TC activity beyond the twenty-first century. Second *seasonal forecasts* of TC activity are issued in December (for the Atlantic) the previous year and are periodically updated throughout the TC season. Last *short-term forecasts* are issued 7–14 days before TC genesis and generally predict intensity and tracks instead of genesis.

For centennial projects, researchers use physics-based climate models. These models project a global decrease in the total number TCs yet are highly uncertain in individual basins. These uncertainties stem from major roadblocks outlined above. First, we don't understand relationships that lead to cyclogenesis (Knutson and Tuleya 2004; LaRow et al. 2008; Zhao et al. 2009) or the climate-TC feedback relationship (Pielke et al. 2005; Emanuel 2008). Second, that the observations are too coarse (20–120 km) to fully model TC properties (Chen et al. 2007).

Seasonal basin-wide activity predictions of seasonal TC activity are issued as early as April in the previous year (to forecast activity in

August–October the following year). Forecasts are generated by both dynamical and statistical models. Similar to physics-based models, dynamical models predict the state of future climate, and the response of the TC-like vortices in the models is used to estimate future hurricane activity (Vitart 2006; Vitart et al. 2007; Vecchi et al. 2011; Zhao et al. 2009). An approach analogous to model simulations is the statistical approach, where one infers relationships solely based on observational data (Elsner and Jagger 2006; Klotzbach and Gray 2009; Gray 1984). These models have limitations based on TC data characteristics. One limitation is that given the relatively short record of observational data, statistical models are subject to overfitting. Another limitation is that the relatively few events make it difficult to interpret a model's output. For instance, if a model predicts a below average season, all it takes is a single strong hurricane to inflict major damage, therefore rendering the forecast uninformative. This was the case in 1983, when hurricane Alicia struck land during a below average season (Elsner et al. 1998). It is no surprise then that these models have yet to impact climate science (Camargo et al. 2007).

Short-term forecasts are used mainly by weather services but have also received attention in the scientific literature. For dynamical models, Belanger et al. (Belanger et al. 2010) test the European Center for Medium-Range Weather Forecasts (ECMWF) Monthly Forecast System's (ECMFS) ability to predict Atlantic TC activity. For the 2008 and 2009 seasons, the model was able to forecast TCs for a week in advance with skill above climatology for the Gulf of Mexico and the MDR on intraseasonal time scales.

Conclusion

We see that forecasting is constrained by characteristics of the data. These constraints provide fertile ground for the ambitious researcher. For instance, we do have good evidence about TC forecasts in the North Atlantic (Pachauri et al. 2014). This suggests that one potential route out

of this bind is to focus on small spatial windows, rather than large basins. Rather than trying to identify relationships in an over parameterized, highly nonlinear, and autocorrelated environment, a focus on smaller manageable windows may generate insights that can be scaled up.

Further Reading

- Belanger, J. I., Curry, J. A., & Webster, P. J. (2010). Predictability of north Atlantic tropical cyclone activity on intraseasonal time scales. *Monthly Weather Review*, 138(12), 4362–4374.
- Camargo, S. J., Barnston, A. G., Klotzbach, P. J., & Landsea, C. W. (2007). Seasonal tropical cyclone forecasts. *WMO Bulletin*, 56(4), 297.
- Chang, E. K. M., & Guo, Y. (2007). Is the number of north Atlantic tropical cyclones significantly underestimated prior to the availability of satellite observations? *Geophysical Research Letters*, 34(14). L14801.
- Chen, S. S., Zhao, W., Donelan, M. A., Price, J. F., & Walsh, E. J. (2007). The CBLAST-hurricane program and the next-generation fully coupled atmosphere–wave–ocean models for hurricane research and prediction. *Bulletin of the American Meteorological Society*, 88(3), 311–317.
- Ding, Y. H., & Reiter, E. R. (1981). *Large-scale circulation conditions affecting the variability in the frequency of tropical cyclone formation over the North Atlantic and the North Pacific Oceans*. Fort Collins, CO: Colorado State University.
- Elsner, J. B., & Jagger, T. H. (2006). Prediction models for annual U.S. hurricane counts. *Journal of Climate*, 19(12), 2935–2952.
- Elsner, J. B., Niu, X., & Tsonis, A. A. (1998). Multi-year prediction model of north Atlantic hurricane activity. *Meteorology and Atmospheric Physics*, 68(1), 43–51.
- Emanuel, K. (2008). The hurricane-climate connection. *Bulletin of the American Meteorological Society*, 89(5), ES10–ES20.
- Gray, W. M. (1984). Atlantic seasonal hurricane frequency. Part I: El Niño's and 30 mb quasi-biennial oscillation influences. *Monthly Weather Review*, 112(9), 1649–1668.
- Gray, W. M., & Brody, L. (1967). *Global view of the origin of tropical disturbances and storms*. Fort Collins, CO: Colorado State University, Department of Atmospheric Science.
- Guha-Sapir, D., Below, R., & Hoyois, P. (2015). *Em-dat: International disaster database*. Brussels: Catholic University of Louvain.
- Hsiang, S. M., & Jina, A. S. (2014). The causal effect of environmental catastrophe on long-run economic

- growth: Evidence from 6,700 cyclones. Technical report, National Bureau of Economic Research.
- Klotzbach, P. J., & Gray, W. M. (2009). Twenty-five years of Atlantic basin seasonal hurricane forecasts (1984–2008). *Geophysical Research Letters* 36(9). L09711.
- Knutson, T. R., & Tuleya, R. E. (2004). Impact of CO₂-induced warming on simulated hurricane intensity and precipitation: Sensitivity to the choice of climate model and convective parameterization. *Journal of Climate*, 17(18), 3477–3495.
- Knutson, T. R., McBride, J. L., Chan, J., Emanuel, K., Holland, G., Landsea, C., Held, I., Kossin, J. P., Srivastava, A. K., & Sugi, M. (2010). Tropical cyclones and climate change. *Nature Geoscience*, 3(3), 157–163.
- Landsea, C. (2007). Counting Atlantic tropical cyclones back to 1900. *Eos, Transactions American Geophysical Union*, 88(18), 197–202.
- LaRow, T. E., Lim, Y.-K., Shin, D. W., Chassignet, E. P., & Cocks, S. (2008). Atlantic basin seasonal hurricane simulations. *Journal of Climate*, 21(13), 3191–3206.
- Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., & Dasgupta, P., et al. (2014). *Climate change 2014: Synthesis report. Contribution of working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC.
- Pielke, R. A., Jr., Landsea, C., Mayfield, M., Laver, J., & Pasch, R. (2005). Hurricanes and global warming. *Bulletin of the American Meteorological Society*, 86(11), 1571–1575.
- Schreck, C. J., III, Knapp, K. R., & Kossin, J. P. (2014). The impact of best track discrepancies on global tropical cyclone climatologies using IBTrACS. *Monthly Weather Review*, 142(10), 3881–3899.
- Vecchi, G. A., Zhao, M., Wang, H., Villarini, G., Rosati, A., Kumar, A., Held, I. M., & Gudgel, R. (2011). Statistical-dynamical predictions of seasonal north Atlantic hurricane activity. *Monthly Weather Review*, 139(4), 1070–1082.
- Vitart, F. (2006). Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Quarterly Journal of the Royal Meteorological Society*, 132(615), 647–666.
- Vitart, F., Huddleston, M. R., Déqué, M., Peake, D., Palmer, T. N., Stockdale, T. N., Davey, M. K., Ineson, S., & Weisheimer, A. (2007). Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophysical Research Letters*, 34(16). L16815.
- Zhao, M., Held, I. M., Lin, S.-J., & Vecchi, G. A. (2009). Simulations of global hurricane climatology, inter-annual variability, and response to global warming using a 50-km resolution GCM. *Journal of Climate*, 22(24), 6653–6678.

Climate Change, Rising Temperatures

Elmira Jamei¹, Mehdi Seyedmohammadian² and Alex Stojcevski²

¹College of Engineering and Science, Victoria University, Melbourne, VIC, Australia

²School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, VIC, Australia

Climate Change and Big Data

In its broadest sense, the term *climate* refers to a statistical description and condition of weather, oceans, land surfaces, and glaciers (considering average and extremes). Therefore, climate change is the alteration in the climate pattern over a long period of time due to both natural and human-induced activities.

The climate of the earth has changed over the past century. Global warming and increased air temperature have significantly altered the ocean, atmospheric condition, sea level, and glaciers. Global climate change, particularly its impact on lifestyle and public health, has become one of the greatest challenges of our era. Human activities and rapid urbanization are known as the main contributors of greenhouse gas emissions. The first scientific assessment of climate change, which was published in June 1990, is a proof for such claim (Houghton et al. 1990). The report is a comprehensive statement on the scientific and climatic knowledge regarding the state of climate change and the role of mankind in exacerbating global warming (Intergovernmental Panel on Climate Change 2015).

To address this rapidly changing climate, there is an urgency to monitor the climate condition, forecast its behavior and identify the most efficient adaptation and mitigation strategies against global warming. This need has already resulted in fruitful outcomes in certain fields, such as science, information technology, and participatory urban

planning. However, despite the urgency of data in climatology, the number of studies highlighting this necessity is lacking.

At present, the amount of climatic data collected rapidly increases. As the volume of climate data increase, data collection, representation, and implementation in decision-making have become equally important. One of the main challenges with increased amount of climatic data lie in information and knowledge management of the collected data on an hourly or even second basis (Flowers 2013).

Big data analytics is one of the methods that help in data monitoring, modeling, and interpretation, which are necessary to better understand causes and effects of climate change and formulate appropriate adaptation strategies. Big data refers to different aspects of data, including data size (such as massive, rapid, or complex) and technological requirement for data processing and analysis.

Big data has had a great success in various fields, such as advertisement and electronic commerce; however, big data is still less employed in climate science. In the era of explosive increasing global data, big data is used to explain and present the massive datasets. In contrast to conventional traditional datasets, big data requires real-time types of analysis. In addition, big data assists in exploring new opportunities and values and achieving an in-depth understanding of hidden values. Big data also addresses a few major questions regarding effective dataset organization and management (Chen et al. 2014).

Exploratory data analysis is the first step prior to releasing data. This analysis is critical in understanding data variabilities and intricacies and is particularly more important in areas such as climate science, where data scientists should be aware regarding the collection process.

Four different sources of climate data include on-site measurements, remote sensing, modeling, and paleoclimate. Each source has its set of strengths and weaknesses which should be fully understood before any data exploration. Table 1 presents each data source with its key strengths and weaknesses.

Climate Change, Rising Temperatures, Table 1
Advantages and disadvantages of climatic data sources (Faghmous and Kumar 2014)

Climate data source	Main strength	Main drawback
Climatic modeling	Capacity to run forward simulations	Only based on physics
On-site measurements and observations	Only based on direct observations	Possibility of having spatial bias
Satellite	Large coverage	Unstable and only lasts for the duration of the mission
Paleoclimate	Capability of using proxy data to infer preindustrial climate trends	Technologies for analyzing such data are still under development

Dealing with continuously changing observation systems is another challenge encountered by climatologists. Data monitoring instruments, especially for satellites and other remote sensing tools, undergo alterations. The change in instrumentations and data processing algorithms poses a question regarding the applicability of such data.

Availability of climatic data before data exploration is another barrier for climatologists. A few datasets have been developed only a decade ago or less. These datasets have spatial resolutions but short temporal duration.

Another barrier in climatic data collection is data heterogeneity. Climate is governed by several interacting variables defined by the earth's system. These variables are monitored and measured using various methods and techniques. However, a few variables cannot be totally observed. For example, a few climatic variables may rely on ground stations; therefore, these variables may be influenced by spatial bias. Other variables may be obtained from satellites whose missions last only 5 or 10 years; thus, continuous monitoring and long recording time is difficult. Despite the fact that the climatic data are sourced from different sources, they belong to a same system and are inter-related. As a result, merging data

from heterogeneous sources is necessary but redundant.

Data representation is another important task for climatologists. Conventional data science is based on attribute–value data. However, certain climatic phenomena (e.g., hurricanes) cannot be represented in the form of attribute–value. For example, hurricanes have their own special patterns; thus, they cannot be represented with binary values. Such evolutionary phenomenon can be demonstrated through equations used in climate models. However, there is still a significant need for similar abstractions within the broader data science.

Conclusion

The rapid acceleration of climate change and global warming are the most significant challenges of the twenty-first century. Thus, innovative and effective solutions are urgently needed. Understanding the changing world and finding adaptation and mitigation strategies have forced researchers with different backgrounds to join together to overcome such issues through a global “data revolution” known as big data. Big data effectively supports climate change research communities in addressing collection, analysis, and dissemination of massive amounts of data and information to enlighten possible future climates under different scenarios, address major challenges encountered by climatologists, and provide guidance to governments in making future decisions.

Further Reading

- Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171–209. <https://doi.org/10.1007/s11036-013-0489-0>.
- Faghmous, J. H., & Kumar, V. (2014). A big data guide to understanding climate change: The case for theory-guided data science. *Big Data*, 2(3), 155–163.
- Flowers, M. (2013). *Beyond open data: The data-driven city. Beyond transparency: Open data and the future of civic innovation* (pp. 185–198). <http://beyondtransparency.org/chapters/part-4/beyond-open-data-the-data-driven-city/>.

- Houghton, J. T., Jenkins, G., & Ephraums, J. (1990). *Climate change: The IPCC scientific assessment. Report prepared for Intergovernmental Panel on Climate Change by working group I*. Cambridge: Cambridge University Press. http://www.ipcc.ch/ipccreports/far/wg_I/ipcc_far_wg_I_full_report.pdf. Accessed 11 June 2012.

- Intergovernmental Panel on Climate Change. (2014). *Climate change 2014: Mitigation of climate change* (Vol. 3). Cambridge University Press.

Cloud

- [Data Center](#)

Cloud Computing

Erik W. Kuiler

George Mason University, Arlington, VA, USA

Cloud-based computing provides important tools for big dataset analytics and management. The cloud-based computing model is a network-based distributed delivery model for providing virtual, on-demand computing services to customers. Cloud-based applications usually operate on multiple Internet-connected computers and servers that are accessible not only via machine-to-machine interactions but also via personal devices, such as smart phones and web browsers. Cloud-based computing is customer focused, offering information technology (IT) capabilities as subscription-based services that require minimal user-direct oversight and management.

Although, in actuality, cloud-based computing may not be the safest option for sensitive data, it may be described via various advantages: no geographical restrictions, cost-effectiveness, reliability, scalability to reflect customers’ needs, and minimal direct requirements for customer-provided active management of cloud-based resources. Additional features include *user-initiated self-service* (on-demand access to network-enabled data storage, server time, applications,

etc.); *network access* (computing capabilities available via a network for use on heterogeneous thick or thin client platforms, e.g., mobile phones, laptops, workstations, etc.); *economies of scale* (resources are pooled and dynamically allocated to meet the demands of multiple customers); *service flexibility* (computer resources provisioned and released to meet customer needs from individual customer perspectives, providing the illusion of access to unlimited resources); and *utilization-based resource management* (resource consumption monitored, measured, and reported to customers and providers of the services).

Infrastructure Implementation Models

Cloud computing configurations of such resources as networks, servers, storage, applications, and services collectively provide enhanced user access to those resources. Cloud infrastructure implementations generally are categorized in relation to four different forms:

1. **Private cloud** – the cloud infrastructure is dedicated to a single organization that may include multiple customers.
2. **Community cloud** – the cloud infrastructure is dedicated to a community of organizations, each of which may have customers that frequently share common requirements, such as security, legal compliance requirements, and missions.
3. **Public cloud** – the cloud infrastructure is open to the public.
4. **Hybrid cloud** – a composition of two or more distinct cloud infrastructures (private, community, or public).

In regard to implementation, cloud-based computing supports different pay-for-use service options. These options include, for example, *Software as a Service (SaaS)* applications that are available by subscription; *Platform as a Service (PaaS)* by which cloud computing providers deploy the cloud infrastructure on which customers can develop and run their own applications; and *Infrastructure as a Service (IaaS)* based on virtual servers, networks, operating

systems, applications, and data storage drives. Note that IaaS is usually an outsourced pay-for-use service; the user usually does not control the underlying cloud structure but may control operating systems, storage, and deployed applications.

Conclusion

Cloud-based computing offers customers a cost-effective, generally reliable means to access and use pooled computer resources on demand, with minimal direct management of those resources. These resources are Internet based and may be geographically dispersed.

Further Reading

- Buyya, R., & Vecchiola, C. (2013). *Mastering cloud computing*. Burlington: Morgan Kaufmann.
- Ji, C., Li, Y., Qiu, W., Awada, U., & Li, K. (2012). Big data processing in cloud computing environments. In *IEEE international symposium on pervasive systems, algorithms and networks*.
- Mell, P., & Grance, T. (2011). NIST special publication 800-145. Available from <https://csrc.nist.gov/publications/detail/sp/800-145/final>.

Cloud Services

Paula K. Baldwin

Department of Communication Studies, Western Oregon University, Monmouth, OR, USA

As consumers and institutions congregate larger and larger portions of data, hardware storage has become inadequate. These additional storage needs led to the development of virtual data centers, also known as the cloud, cloud computing, or, in the case of the cloud providers, cloud services. The origin of the term, cloud computing, is somewhat unclear, but a cloud-shaped symbol is often used as a representation on the Internet of the cloud. The cloud symbol also represents the remote, complex system infrastructure used to store and manage the consumer's data.

The first reference to cloud computing in the contemporary age appeared in the mid-1990s, and it became popular in the mid-2000s. As cloud services become much more versatile and economical, consumers' use is increasing. The cloud offers users immediate access to a shared pool of computer resources. As processors continue to develop both in power and economic feasibility, the expansion of these data centers (the cloud) has expanded on an enormous scale. Cloud services incentivize migration to the cloud as users recognize the elastic potential for data storage as a reasonable cost. Cloud services are the new generation of computing infrastructures, and there are multiple cloud vendors providing a range of cloud services. The fiscal benefit of cloud computing is the consumer only pays for the use on the resources they need without any concern over compromising their physical storage areas. The cloud service manages the data on the back end. In an era where physical storage limitations has become problematic with increased downloads of movies, books, graphics, and other high data memory products, cloud computing has been a welcome development.

Choosing a Cloud Service

As the cloud service industry grows, choosing a cloud service can be confusing for the consumer. One of the first areas to consider is the unique cloud service configurations. Cloud services are configured in four ways. One, public clouds may be free or bundled with other services or offered as pay per usage. Generally speaking, public cloud service providers like Amazon AWS, Microsoft, and Google own and operate their own infrastructure data centers, and access to these providers' services is through the Internet. Private cloud services are data management infrastructures created solely for one particular organization. Management of the private cloud may be internal or external. Community cloud services exist when multiple organizations from a specific community with common needs choose to share an infrastructure. Again, management of the community cloud service may be internal or external, and fiscal

responsibility is shared between the organizations. Hybrid clouds are a grouping of two or more clouds, public or private community, where the cloud service is comprised of variant combination that extends the capacity of the service through aggregation, integration, or customizations with another cloud service. Sometimes a hybrid cloud is used on a temporary basis to meet short-term data needs that cannot be fulfilled by the private cloud. Having the ability to use the hybrid cloud enables the organization to only pay for the extra resources when they are needed, so this exists as a fiscal incentive for organizations to use a hybrid cloud service.

The other aspect to consider when evaluating cloud services is the specific service models offered for the consumer or organization. Cloud computing offers three different levels of service: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). The SaaS has a specific application or service subscription for the customer (e.g., Dropbox, [Salesforce.com](https://www.salesforce.com), and QuickBooks). With the SaaS, the service provider handles the installation, setup, and running of the application with little to no customization. The PaaS allows businesses an integrated platform on which they can create and deploy custom apps, databases, and line-of-business service (e.g., Microsoft Windows Azure, IBM Bluemix, Amazon Web Services (AWS), Elastic Beanstalk, Heroku, [Force.com](https://www.force.com), Apache Stratos, Engine Yard, and Google App Engine). The PaaS service model includes the operating system, programming language execution environment, database, and web server designed for a specific framework with a high level of customization. With Infrastructure as a Service (IaaS), businesses can purchase infrastructure from providers as virtual resources. Components include servers, memory, firewalls, and more, but the organization provides the operating system. IaaS providers include Amazon Elastic Cloud Computer (Amazon EC2), GoGrid, Joyent, AppNexus, Rackspace, and Google Compute Engine.

Once the correct cloud service configuration is determined, the next step is to match user needs with correct service level. When looking at cloud services, it is important to examine four different

aspects: application requirements, business expectations, capacity provisioning, and cloud information collection and process. These four areas complicate the process of selecting a cloud service. First, the application requirements refer to the different features such as data volume, data production rate, data transfer and updating, communication, and computing intensities. These factors are important because the differences in these factors will affect the CPU (central processing unit), memory, storage, and network bandwidth for the user. Business expectations fluctuate depending on the applications and potential users, which, in turn, affect the cost. The pricing model depends on the level of the service required (e.g., voicemail, a dedicated service, amount of storage required, additional software packages, and other custom services). Capacity provisioning is based on the concept that, according to need, different IT technologies are employed and, therefore, each technology has its own unique strengths and weaknesses. The downside for the consumer is the steep learning curve required. The final challenge requires that the consumers invest a substantial amount of time to investigate individual websites, collect information about each cloud service offering, collate their findings, and employ their own assessments to determine their best match. If an organization has an internal IT department or employs an IT consultant, the decision is easier to make; for the individual consumer, without an IT background, the choice may be considerably more difficult.

Cloud Safety and Security

For the consumer, two primary issues are relevant to cloud usage: a check and balance system on the usage versus service level purchased and data safety. This on-demand computation model of cloud computing is processed through large virtual data centers (clouds), offering storage and computation needs for all types of cloud users. These needs are based on service level agreements. Although cloud services are relatively low cost, there is no way to know if the services they are purchasing are equivalent to the service level purchased. Although being able to

determine that a consumer's usage in relationship to the service level purchased is appropriate, the more serious concern for consumers is data safety. Furthermore, because users do not have physical possession of their data, public cloud services are underutilized due to trust issues. Larger organizations use privately held clouds, but if a company does not have the resources to develop their own cloud service, most organizations are unlikely to use public cloud services due to safety concerns. Currently, there is no global standardization of data encryption between cloud services, and there have been some concerns raised by experts who say there is no way to be completely sure that data, once moved to the cloud, remains secure. With most cloud services, control of the encryption keys is retained by the cloud service, making your data vulnerable to a rogue employee or a governmental request to see your data.

The Electronic Frontier Foundation (EFF) is a privacy advocacy group that maintains a section on their website (*Who Has Your Back*) that rates the largest Internet companies on their data protections. The EFF site uses six criteria to rate the companies: requires a warrant for content, tells users about government data requests, publishes transparency reports, publishes law enforcement guidelines, fights for user privacy rights in courts, and fights for user privacy rights in Congress. Another consumer and corporate data protection group is the Tahoe Least Authority File System (Tahoe-LAFS) project. Tahoe-LAFS protects a free, open-source storage system created and developed by Zooko Wilcox-O'Hearn with the goal of data security and protection from hardware failure. The strength of this storage system is their encryption and integrity – checks first go through gateway servers, and after the process is complete, the data is stored on a secondary set of servers that cannot read or modify the data.

Security for data storage via cloud services is a global concern whether for individuals or organizations. From a legal perspective, there is a great deal of variance in how different countries and regions deal with security issues. At this point in time, until there are universal rules or legacy specifically addressing data privacy legislation, the consumers must take responsibility for their own data. There are five strategies for keeping

your data secure in the cloud, outside of what the cloud services offer. First, consider storing crucial information somewhere other than the cloud. For this type of information, perhaps utilizing the available hardware storage might be the best solution rather than using a cloud service. Second, when choosing a cloud service, take the time to read the user agreement. The user agreement should clearly delineate the parameters of their service level and that will help with the decision-making. Third, take creating passwords seriously. Oftentimes, the easy route for passwords is familiar information such as dates of birth, hometowns, and pet's or children's names. With the advances in hardware and software designed specially to crack passwords, it is particularly important to use robust, unique passwords for each of your accounts. Fourth, the best way to protect data is through encryption. The way encryption works in this instance is to use an encryption software on a file before you move the file to the cloud. Without the password to the encryption, no one will be able to read the file content. When considering a cloud service, investigate their encryption services. Some cloud services encrypt and decrypt user files local as well as provide storage and backup. Using this type of service ensures that data is encrypted before it is stored in the cloud and after it is downloaded from the cloud providing the optimal safety net for consumer data.

Cross-References

- [Cloud](#)
- [Cloud Computing](#)
- [Cloud Services](#)

Further Reading

- Ding, S., et al. (2014). Decision support for personalized cloud service selection through multi-attribute trustworthiness evaluation. *PLoS One*, 9(6), e97762.
- Gui, Z., et al. (2014). A service brokering and recommendation mechanism for better selecting cloud services. *PLoS One*, 8(8). e105297. <https://doi.org/10.1371/journal.pone.0105297>.
- Hussain, M., et al. (2014). Software quality in the clouds: A cloud-based solution. *Cluster Computing*, 17(2), 389–402.

- Kun, H., et al. (2014). Securing the cloud storage audit service: Defending against frame and collude attacks of third party auditor. *IET Communications*, 8(12), 2106–2113.
- Mell, P., et al. (2011). National Institute of Standards and Technology, U.S. Department of Commerce. The NIST definition of cloud computing. Special Publication 800-145, 9–17.
- Qi, Q., et al. (2014). Cloud service-aware location update in mobile cloud computing. *IET Communications*, 8(8), 1417–1424.
- Rehman, Z., et al. (2014). Parallel cloud service selection and ranking based on QoS history. *International Journal of Parallel Programming*, 42(5), 820–852.

Cluster Analysis

- [Data Mining](#)

Collaborative Filtering

Ashrf Althbiti and Xiaogang Ma
Department of Computer Science, University of Idaho, Moscow, ID, USA

Synonyms

[Data reduction](#); [Network data](#); [Recommender systems](#)

Introduction

Collaborative filtering (CF) entirely depends on users' contribution such as ratings or reviews about items. It exploits the matrix of collected user-item ratings as the main source of input. It ultimately provides the recommendations as an output that takes the following two forms: (1) a numerical prediction to items that might be liked by an active user U and (2) a list of top-rated items as top- N items. CF claims that similar users express similar patterns of rating behavior. Also, CF claims that similar items obtain similar ratings. There are two primary approaches of CF algorithms: (1) neighborhood-based and (2) model-based (Aggarwal 2016).

The neighborhood-based CF algorithms (aka, memory-based) directly utilize stored user-item ratings to predict ratings for unseen items. There are two primary forms of neighborhood-based algorithms: (1) user-based nearest neighbor CF and (2) item-based nearest neighbor CF (Aggarwal 2016). In the user-based CF, two users are similar if they rate several items in a similar way. Thus, it recommends to a user the items that are the most preferred by similar users. In contrast, the item-based CF recommends to a user the items that are the most similar to the user's previous purchases. In such an approach, two items are similar if several users have rated these items in a similar way.

The model-based CF algorithms (aka, learning-based models) form an alternative approach by sending both items and users to the same latent factor space. The algorithms utilize users' ratings to learn a predictive model (Ning et al. 2015). The latent factor space attempts to interpret ratings by characterizing both items and users on factors automatically inferred from previous users' ratings (Koren and Bell 2015).

Methodology

Neighborhood-Based CF Algorithms

User-Based CF

User-based CF claims that if users rated items in a similar fashion in the past, they will give similar

rating to new items in the future. For instance, Table 1 shows a user-item ratings matrix, which includes four users' rating of four items. The task is to predict the rating of unrated *item3* by the active user *Andy*.

In order to solve the task presented above, the following notations are given. The set of users is symbolized as $U = \{U1, \dots, Uu\}$, the set of items is symbolized as $I = \{I1, \dots, Ii\}$, the matrix of ratings is symbolized as R where $r_{u,i}$ means rating of a user U for an item I , and the set of possible ratings is symbolized as S where its values take a range of numerical ratings $\{1, 2, 3, 4, 5\}$. Most systems consider the value 1 as strongly dislike and the value 5 as strongly like. It is worth noting that $r_{u,i}$ should only take one rating value.

The first step is to compute the similarity between *Andy* and the other three users. In this example, the similarity between the users is simply computed using Pearson's correlation coefficient (1).

$$sim(u, v) = \frac{\sum_{i \in I(ru, i - \bar{ru})(rv, i - \bar{rv})} / \sqrt{\sum_{i \in I(ru, i - \bar{ru})^2} * \sqrt{\sum_{i \in I(rv, i - \bar{rv})^2}}}{\quad} \quad (1)$$

where $\bar{ru}$ and $\bar{rv}$ are the average rating of the available ratings made by users u and v .

By applying Eq. (1) to the rating data in Table 1 given that ($r_{Andy} = \frac{3+3+5}{3} = 3.6$, and $r_{U1} = \frac{4+2+2+4}{4} = 3$), the similarity between *Andy* and *U1* is calculated as follows:

$$sim(Andy, U1) = \frac{(3 - 3.6)(4 - 3) + (3 - 3.6)(2 - 3) + (5 - 3.6)(4 - 3)}{\sqrt{(3 - 3.6)^2 + (3 - 3.6)^2 + (5 - 3.6)^2} \times \sqrt{(4 - 3)^2 + (2 - 3)^2 + (4 - 3)^2}} = 0.49 \quad (2)$$

It is worth noting that the results of Pearson's correlation coefficient are in the range of $(+1$ to $-1)$, where $+1$ means high positive correlation and -1 means high negative correlation. The similarities between *Andy* and *U2* and *U3*

are 0.15 and 0.19, respectively. Referring to the previous calculations, it seems that *U1* and *U3* similarly rated several items in the past. Thus, *U1* and *U3* are utilized in this example to predict the rating of *item3* for *Andy*.

The second step is to compute the prediction for *item3* using the ratings of *Andy's K*-neighbors (*U1* and *U3*). Thus, Eq. (3) is introduced where $\hat{r}$ means the predicted rating.

$$\hat{r}(u, i) = \overline{ru} + \frac{\sum_{v \in K} \text{sim}(u, v) * (rv, i - \overline{rv})}{\sum_{v \in K} \text{sim}(u, v)} \quad (3)$$

$$\begin{aligned} \hat{r}(\text{Andy}, \text{item3}) &= \overline{r\text{Andy}} \\ &+ \frac{\text{sim}(\text{Andy}, \text{U1}) * (r\text{U1}, \text{item3} - \overline{r\text{U1}}) + \text{sim}(\text{Andy}, \text{U3}) * (r\text{U3}, \text{item3} - \overline{r\text{U3}})}{\text{sim}(\text{Andy}, \text{U1}) + \text{sim}(\text{Andy}, \text{U3})} \\ &= 4.45 \end{aligned} \quad (4)$$

Given the result of the prediction computed by Eq. (4), it is most likely that *item3* will be a good choice to be included in the recommendation list for *Andy*.

Item-Based CF

Item-based CF algorithms are introduced to solve serious challenges when applying user-based nearest neighbor CF algorithms. The main challenge is that when the system has massive records of users, the complexity of the prediction task increases sharply. Accordingly, if the number of items is less than the number of users, it is ideal to adopt the item-based CF algorithms.

This approach computes the similarity between items instead of an enormous number of potential neighbor users. Also, this approach considers the ratings of user *U* to make a prediction for item *I*, as item *I* will be similar to the previous rated items by user *U*. Therefore, users may prefer to utilize their ratings rather than other users' rating when making the recommendations.

Equation (5) is used to compute the similarity between two items.

$$\begin{aligned} \text{sim}(i, j) &= \frac{\sum_{u \in U} (ru, i - \overline{ri})(ru, j - \overline{rj})}{\sqrt{\sum_{u \in U} (ru, i - \overline{ri})^2} \sqrt{\sum_{u \in U} (ru, j - \overline{rj})^2}} \\ &\times \sqrt{\sum_{u \in U} (ru, j - \overline{rj})^2} \end{aligned} \quad (5)$$

In Equation (5), $\overline{ri}$ and $\overline{rj}$ are the average rating of the available ratings made by users for both items *i* and *j*.

Then, make the prediction for item *I* for user *U* by applying Eq. (6) where *K* means the number of neighbors of items for item *I*.

$$\begin{aligned} \hat{r}(u, i) &= \overline{ri} \\ &+ \frac{\sum_{j \in K} \text{sim}(i, j) * (ru, j - \overline{rj})}{\sum_{j \in K} \text{sim}(i, j)} \end{aligned} \quad (6)$$

Model-Based CF Algorithms

Model-based CF algorithms take the raw data that has been preprocessed in the offline step where the data typically requires to be cleansed, filtered, and transformed and then generate the learned model to make a prediction. It solves several issues that appear in the neighborhood-based CF algorithms. These issues are (1) limited coverage which means finding neighbors is based on the rating of common items and (2) sparsity in the rating matrix which means the diversity of items rated by different users.

Model-based CF algorithms compute the similarities between users or items by developing a parametric model that investigates their

Collaborative Filtering, Table 1 User-item rating dataset

User name	Item1	Item2	Item3	Item4
Andy	3	3	?	5
U1	4	2	2	4
U2	1	1	4	2
U3	5	2	3	4

relationships and patterns. It is classified into two main categories: (1) factorization methods and (2) adaptive neighborhood learning methods (Ning et al. 2015).

Factorization Methods

Factorization methods aim to define the characterization of ratings by projecting users and items to the reduced latent vector. It helps discover more expressive relations between each pair of users, items, or both. It has two main types: (1) factorization of a sparse similarity matrix and (2) factorization of an actual rating matrix (Jannach et al. 2010).

The factorization is done by using the singular value decomposition (SVD) or the principal component analysis (PCA). The original sparse ratings or similarities matrix is decomposed into a smaller-rank approximation in which it captures the highly correlated relationships. It is worth mentioning that the SVD theorem (Golub and Kahan 1965) claims that matrix M can be collapsed into a product of three matrices as follows:

$$M = U \Sigma V^T \quad (7)$$

where U and V contain left and right singular vectors and the values of the diagonal of Σ are singular values.

Adaptive Neighborhood Learning Methods

This approach combines the original neighborhood-based and model-based CF methods. The main difference of this approach, in comparison with the basic neighborhood-based, is that the learning of the similarities is directly inferred from the user-item ratings matrix, instead of adopting pre-defined neighborhood measures.

Conclusion

This article discusses a general perception of the CF. CF is one of the early approaches proposed for information filtering and recommendation making. However, CF still ranks among the most popular methods that people employ in

nowadays for researches on Web, big data, and data mining.

Cross-References

- [Data Aggregation](#)
- [Data Cleansing](#)
- [Network Analytics](#)

References

- Aggarwal, C. C. (2016). An introduction to recommender systems. In *Recommender systems* (pp. 1–28). Cham: Springer.
- Golub, G., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2), 205–224.
- Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: An introduction*. Cambridge, UK: Cambridge University Press.
- Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. In *Recommender systems handbook* (pp. 77–118). Boston: Springer.
- Ning, X., Desrosiers, C., & Karypis, G. (2015). A comprehensive survey of neighborhood-based recommendation methods. In *Recommender systems handbook* (pp. 37–76). Boston: Springer.

Column-Based Database

- [NoSQL \(Not Structured Query Language\)](#)

Common Sense Media

Dzmitry Yuran

School of Arts and Communication, Florida
Institute of Technology, Melbourne, FL, USA

The rise of big data has brought us to the verge of redefining our understanding of privacy. The possibility of high-tech profiling, identification, and discriminatory treatment based on information often provided unknowingly (and

sometimes involuntarily) brings us to the forefront of a new dimension of the use and protection of personal information. Even less aware than adults of the means to the ends of their digital product consumption, children become more vulnerable to the risks of the digital world defined by big data.

The issue of children's online privacy and protection of their safety and rights in today's virtually uncontrolled Internet environment is among the main concerns of Common Sense Media (CSM), an independent non-profit organization providing parents, educators, and policymakers with tools and advice to aid making children's use of media and technology a safer and more positive experience.

Protecting data that students and parents provide to education institutions from commercial interest and other third parties is the key concern behind CSM's School Privacy Zone campaign. The organization does more to advocate safeguarding the use of media by youth. It also provides media ratings and reviews, designs, and promotes educational tools.

Media Reviews and Ratings

On their website, www.common sense media.org, Common Sense Media publishes independent expert reviews of movies, games, television programs, books, application software, websites and music. The website enables sorting through the reviews by media type, age, learning rating, topic, genre, required hardware and software platforms, skills the media are aimed at developing or improving, recommendations by editors, parents, as well as popularity among children.

Common Sense Media does not accept payments for its reviews in order to avoid bias and influence by creators and publishers (charitable donations are welcome, however). Media content is reviewed and rated by a diverse trained staff (from reviewers for major publications to librarians, teachers, and academics) and edited by a group of writers and media professionals.

All reviewed media content is assigned one of the four age-appropriateness ratings, depending

on the selected age group: ON (appropriate), PAUSE (some content could be suitable for some children of a selected age group), OFF (not age-appropriate), and NOT FOR KIDS (inappropriate for kids of any age). A calculated score from a series of five-point scale categories (such as "positive role models," "positive messages," "violence & scariness," etc.) determines the assignment of an ON, PAUSE, or OFF rating to content for an age group.

Some media, such as software applications, video games, and websites, are evaluated according to their learning potential on a five-point scale (three-point scale before 2013), ranging from BEST (excellent learning approach) to NOT FOR KIDS (not recommended for learning). Each rated item receives a series of scores in dimensions such as engagement, learning approach, feedback, and other. Combined score across the dimensions determines overall learning potential of given media content.

A one to five star rating assesses media's overall quality. Parents, children, and educators can review and rate media after creating an account on the Common Sense Media website. User reviews and rating are displayed separately and are broken down into two groups: parents and kids reviews.

As of summer 2014, the Common Sense Media review library was exceeding 20,000 reviews.

Education

Common Sense Media provide media and technology resources for educators, including Graphite, a free tool for discovery and sharing of curricula, educational software, sites and games. As a part of Graphite, App Flows, an interactive lesson plan-building framework allows educators to fit discovered digital tools into a dynamic platform to create and share lesson plans.

Common Sense also hosts Appy Hours, a series of videos on the organization's YouTube channel, designed to bring educators together for a discussion of ways in which digital tools could be used for learning.

Editorial picks for educational digital content, discussion boards, and blogs are incorporated into

the Common Sense Graphite site in order to enhance educators' experience with the system.

Advocacy

Common Sense Media works with lawmakers and policymakers nationwide in the pursuit of an improved media landscape for children and families. The main issues the organization addresses are represented in three areas: children's online privacy, access to digital learning, violence and gender roles in media.

In an attempt to give more control over kids' digital footprints to children themselves as well as their families, CSM supports several legislative projects, including the Do Not Track Kids bill concerned with collecting location information and sending targeted ads to teens as well as the Eraser Button bill which requires apps and websites to allow teens to remove their postings and prohibits advertisement of illegal or harmful products, like alcohol or tobacco, to minors. The organization also promotes the need for updates to the Federal Trade Commission's 1999 Children's Online Privacy Protection Act designed to give parents more control over information about their children collected and shared by companies online.

In their effort to promote digital learning technology, Common Sense Media supports Federal Communication Commission's E-Rate program aimed at bringing high-speed Internet to American schools. The CSM's School Privacy Zone initiative seeks to safeguard data about students collected by schools and educators from advertisers and other private interests.

The organization addresses the concern with the impact that video games and other media content could have on development of children, as well as the contribution of these media to the culture of violence in the United States. CSM highlights the gaps in research of portrayal of violence in media and encourages Congress to promote further scientific inquiry in order to address overwhelming concern among parents.

Research

Common Sense Media carries out a variety of research projects in order to inform its ratings and aid its advocacy efforts. They collect, analyze, and disseminate data on children's use of media and technology and media impact on their development. Both original data (collected via commissioned by Common Sense Media to Knowledge Networks online surveys) and secondary data from large national surveys and databases are used in their analyses. The organization also produces overviews of the state of research on certain topics, such as advertising for children and teens and media connection to violence. Full texts of featured reports as well as summaries and infographics highlighting main findings are available for viewing and download on commonsensemedia.org free of charge.

Results of Commons Sense Media research regularly make their way into mainstream mass media. Among others, NPR, Time, and the New York Times featured CSM findings in news stories.

Organization History, Structure and Partnerships

Common Sense Media was founded in 2003 by James P. Steyer, the founder of Children Now Group and a lecturer at Stanford University at the time. With initial investment of \$500,000 from various backers (including among others Charles R. Schwab of the Charles Schwab Corporation, Philip F. Anschutz of Qwest Communications International, George R. Roberts of Kohlberg Kravis Roberts & Company, and James G. Coulter of Texas Pacific Group) the first office opened in San Francisco, CA. William E. Kennard and Newton N. Minow, two former Federal Communications Commission chairmen, were among the first board members for the young organization.

Since 2003, Common Sense Media has grown to 25 members on the board of directors and

25 advisors. It employs extensive teams of reviewers and editors. The organization added three regional offices: in Los Angeles, New York City, and Washington D.C. and established a presence in social media (Facebook, Twitter, YouTube, Google+, and Pinterest).

The evolution of the Internet and the ever-growing number of its applications turned it into a virtual world with almost endless possibilities. While the rules and the laws of this new world are yet to take shape and be recorded, young people spend more and more time in this virtual reality. It affects their lives both inside and outside the virtual world, affecting their development, and physical and emotional state. Virtually any activity we engage in online produces data which could be used in both social research and with commercial purposes. Collection and use of these data (for advertising or other purposes) could pose serious legal and ethical issues which raises serious concern among parents of young media consumers and educators. As some of the most vulnerable media consumers, children need additional protection and guidance in the virtual world. Organizations like Common Sense Media, parents, educators, law makers, and policy makers begin paying closer attention to the kids' place on virtual reality and the impact that reality can have on children.

Cross-References

- [Media](#)
- [Online Advertising](#)
- [Social Media](#)

Further Reading

- Common Sense Media. Graphite™: About us <http://www.graphite.org/about-us>. Accessed Sept 2014.
- Common Sense Media. Our mission. <https://www.commonsensemedia.org/about-us/our-mission#about-us>. Accessed Sept 2014.
- Common Sense Media. Policy priorities. <https://www.commonsensemedia.org/advocacy>. Accessed Sept 2014.

Common Sense Media. Program for the study of children and media. <https://www.commonsensemedia.org/research>. Accessed Sept 2014.

Rutenberg, J. (2003). A new attempt to monitor media content. *The New York Times*. <http://www.nytimes.com/2003/05/21/business/a-new-attempt-to-monitor-media-content.html>. Accessed Sept 2014.

Communication Quantity

Martin Hilbert

Department of Communication,

University of California, Davis, Davis, CA, USA

An increasing share of the world's data capacity is centralized in "the cloud." The gatekeeper to obtain access to this centralized capacity is telecommunication access. Telecommunication channels are the necessary (but not sufficient) condition to provide access to the mass of the world's data storage.

In this inventory we mainly follow the methodology of what has become a standard reference in estimating the world's technological information capacity: Hilbert and López (2011). The total communication capacity is calculated as the sum of the product of technological devices and their bandwidth performance, where the latter is normalized on compression rates. We measure the "installed capacity" (not the effectively used capacity), which implies that it is assumed that all technological capacities are used to their maximum. For telecommunication, this describes the "end-user bandwidth potential" ("if all end-users would use their full bandwidth"). This is merely a "potential," because in reality, negative network externalities create a trade-off in bandwidth among users. For example, estimating that the average broadband connection is 10 Mbps in a given country does not mean that all users could use this average bandwidth at the same second. The network would collapse. The normalization on software compression rates is important for the creation of meaningful time series, as

compression algorithms have enable to send more information through the same hardware infrastructure over recent decades (Hilbert 2014a; Hilbert and López 2012a). We normalize on “optimally compressed bits” as if all content were compressed with the best compression algorithms possible in 2014 (Hilbert and López 2012b). For the estimation of compression rates of different content, justifiable estimates are elaborated for 7-year intervals (1986, 1993, 2000, 2007, 2014). The subscriptions data stem mainly from ITU (2015) with completions from other sources. One of the main sources for internet bandwidth is NetIndex (Ookla 2014), which has gathered the results of end-user-initiated bandwidth velocity tests per country per day over recent years (e.g., an average 180,000 test per day already in 2010 through Speedtest.net and Pingtest.net). For more see Hilbert (2015) and López and Hilbert (2012).

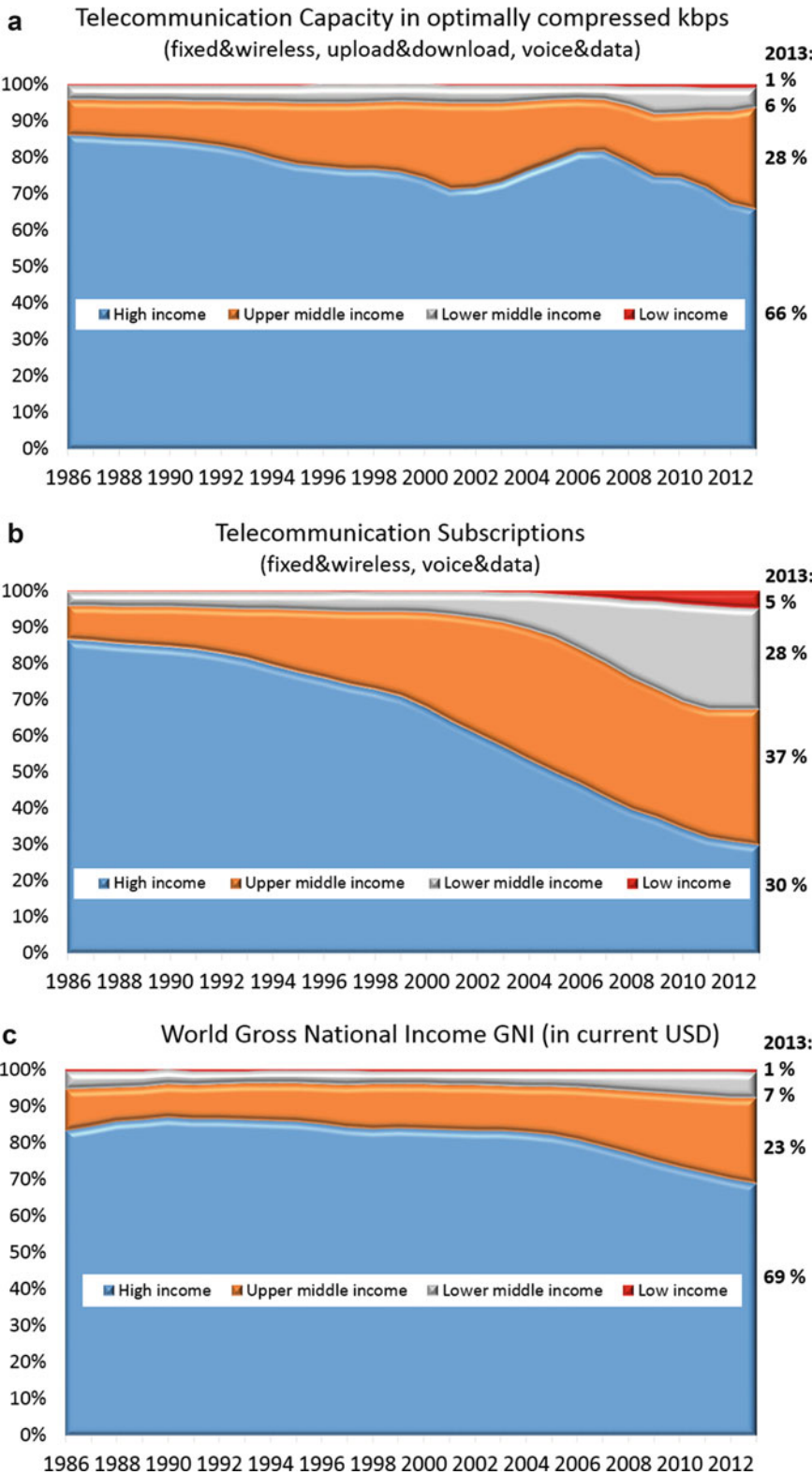
Figure 1a looks at the total telecommunication capacity in optimally compressed kbps in terms of global income groups (following the classification of the World Bank of 2015). The world’s installed telecommunication capacity has grown with a compound annual growth rate of 35% during the same period, (from 7.5 petabits to 25 exabits). The last three decades show a gradual loss of dominance of global information capacities for today’s high-income countries. High-income countries dominated 86% of the globally installed bandwidth potential, but merely 66% in 2013. It is interesting to compare this presentation with the more common method to assess the advancement approximation in terms of the number of telecommunication subscriptions (Fig. 1b). Both dynamics are quite different, which stems from the simple fact that not all subscriptions are equal in their communicational performance. This intuitive difference is the main reason why the statistical accounting of subscriptions is an obsolete and very often misleading indicator. This holds especially true in an age of Big Data, where the focus of development is set on informational bits, not on the number of technological devices (for the complete argument, see Hilbert (2014b, 2016)).

Comparing these results with the global shares of Gross National Income (GNI) and population (Fig. 1c, d), it becomes clear that the diffusion

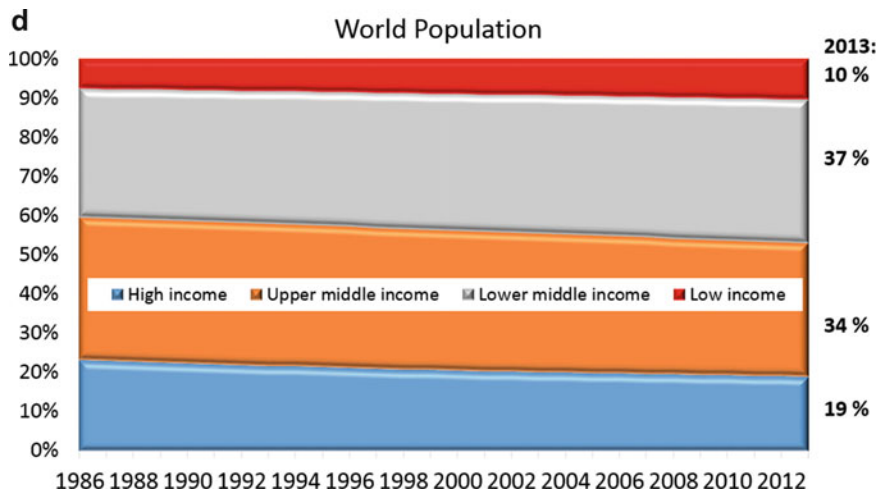
dynamic of the number of subscriptions follows existing patterns in population distribution. Especially the diffusion of mobile phones during recent decades has contributed to the fact that both distributions align. The number of subscriptions reaches a saturation limit at about 2–2.5 subscriptions per capita worldwide, and therefore leads to a natural closure of the divide over time. On the contrary, communication capacity in kbps (and therefore access to the global Big Data infrastructure) follows the signature of economic capacities. After only a few decades, both processes align impressively well. This shows that the digital divide in terms of data capacity is far from being closed but is rather becoming a structural characteristic of modern societies, which is as persistent as the existing income divide (Hilbert 2014b, 2016).

Figure 1a also reveals that the evolution of communication capacities in kbps is not a monotone process. Increasing and decreasing shares between high income and upper middle income countries suggest that the evolution of bandwidth is characterized by a complex nonlinear interplay of public policy, private investments, and technological progress. Some countries in this income range seem to (at least temporarily) do much better than their economic capacity would suggest. This is a typical signature of effective public policy.

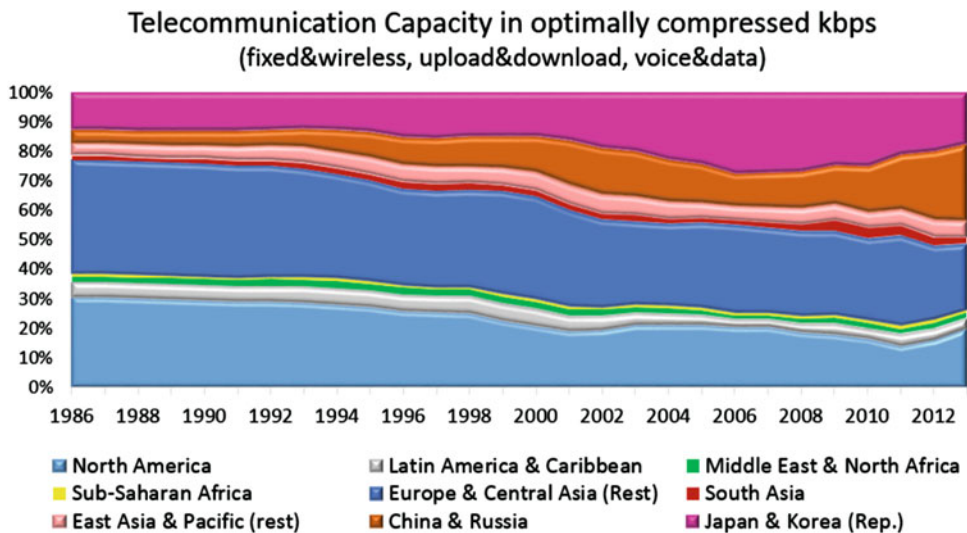
Figure 2 shows the same global capacity in optimally compressed kbps per geographic regions (following the World Bank classification of 2015). Asia has notably increased its global share at the expense of North America and Europe, with a share of less than a quarter of the global capacity in 1986 (23%) and a global majority of 51% in 2013 (red-shaded areas in Fig. 2). Figure 2 reveals that the main driver of this expansion during the early 2000s were Japan and South Korea, both of which famously pursued a very aggressive public sector policy agenda in the expansion of fiber optic infrastructure in the early 2000s. The more recent period since 2010 is characterized by the expansion of bandwidth in both China and Russia. Notably, most recent broadband policy efforts in the USA seems to show some first detectable effects on a macrolevel, as North America has started to return its tendency of a shrinking global share during recent years.



Communication Quantity, Fig. 1 (continued)



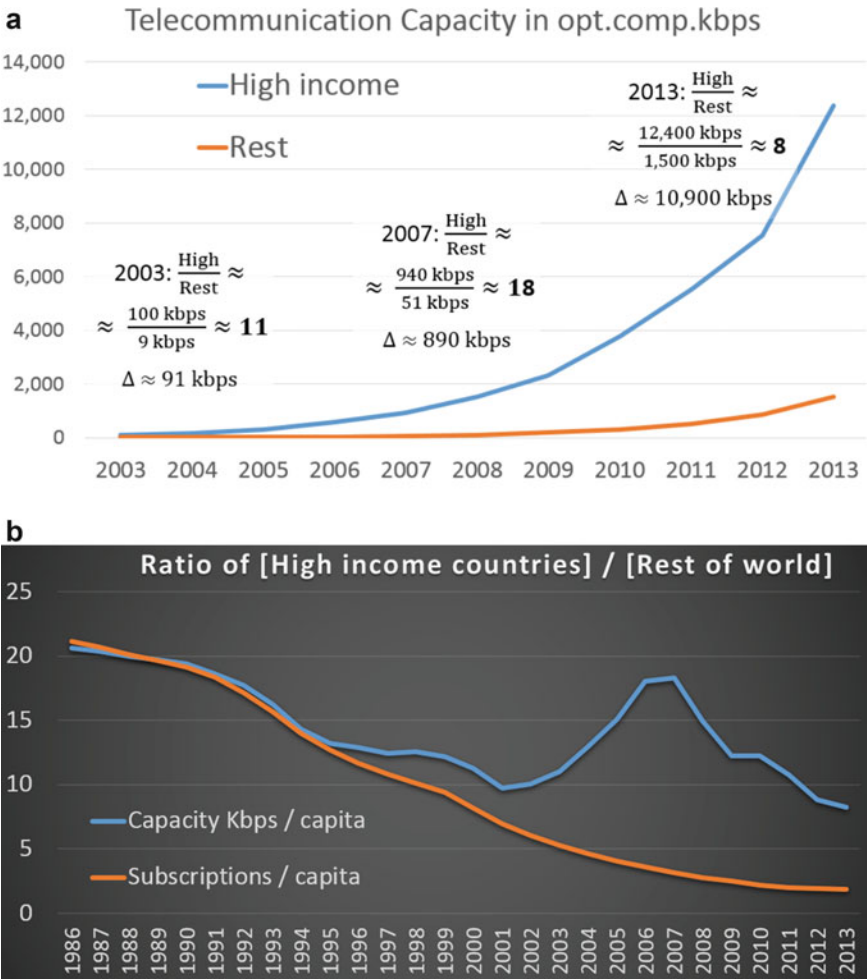
Communication Quantity, Fig. 1 International income groups: (a) telecommunication capacity in optimally compressed kbps; (b) telecommunication subscriptions; (c) World Gross National Income (GNI, current USD); (d) World population



Communication Quantity, Fig. 2 Telecommunication capacity in optimally compressed kbps per world region

Expressed in installed kbps per capita (per inhabitant), we can obtain a clearer picture about the increasing and decreasing nature of the evolving digital divide in terms of bandwidth capacity. First and foremost, Fig. 3a shows that the divide continuously increases in absolute terms. In 2003, the average inhabitant of high-income countries

had access to an average of 100 kbps of installed bandwidth potential, while the average inhabitant of the rest of the world had access to merely 9 kbps. In absolute terms, this results in a difference of some 90 kbps. As shown in Fig. 3a, this divide increased with an order of magnitude every 5 years, reaching almost 900 kbps in 2007 and



Communication Quantity, Fig. 3 (a) Telecommunication capacity per capita in optimally compressed kbps: high-income groups (World Bank classification) versus

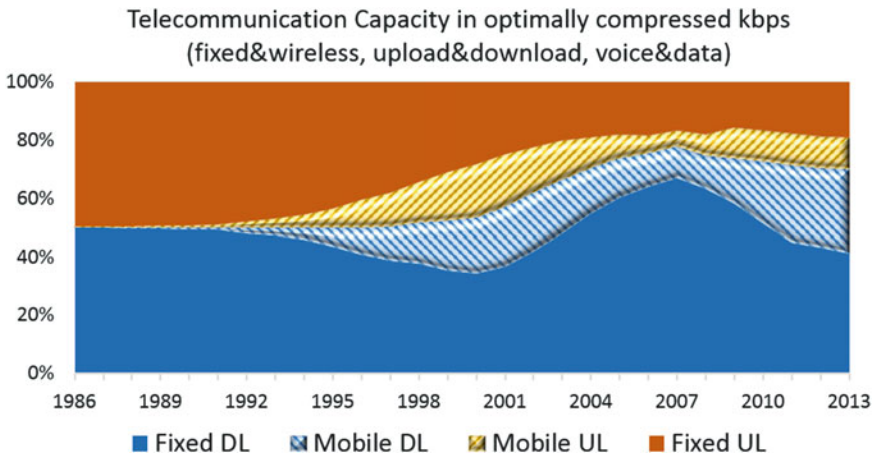
rest of world. (b) Ratio of telecommunication capacity per capita in high-income countries versus rest of world, and of subscriptions per capita

over 10,000 kbps by 2013. This increasing divide in absolute terms is important to notice in the context of a Big Data world, in which the amount of data is becoming a crucial ingredient for growth.

In relative terms, this results in an increasing and decreasing evolution of the divide over time. Figure 3b contrasts this tendency with the monotonically decreasing tendency of the digital divide in terms of telecommunication subscriptions. It shows that the divide in terms of data capacities is much more susceptible to both technological change and technology interventions. The decreasing divide during the period until 2000 is explained

by the global diffusion of narrowband internet and 2G telephony. The increasing nature of the divide between 2001 and 2008 is due to the global introduction of broadband for fixed and mobile solutions. The most recent decreasing nature of the divide is evidence of the global diffusion of broadband. The digital divide in terms of data capacities is a continuously moving target, which opens up with each new innovation that is introduced into the market (Hilbert 2014b, 2016).

Finally, another aspect with important implications for the Big Data paradigm is the relation between uplink and downlink capacity. Uplink



Communication Quantity, Fig. 4 Telecommunication capacity in optimally compressed kbps per uplink and downlink

and downlinks show the potential of contribution and exploitation of the digital Big Data footprint. Figure 4 shows that the global telecommunication landscape has evolved from being a media of equal up- and downlink, toward to more download heavy medium. Up until 1997, global telecommunication bandwidth potential was equally split with 50% up- and 50% down-link. The introduction of broadband and the gradual introduction of multimedia video and audio content changed this. In 2007, the installed uplink potential was as little as 22%. The global diffusion of fiber optic cables seems to reverse this trend, reaching a share of 30% uplink in 2013. It can be expected that the share of effectively transmitted bits through this installed bandwidth potential leads to an even larger share of fixed-line broadband (for more in these methodological differences, see Hilbert and López (2012a, b)).

Further Reading

- Hilbert, M. (2014a). How much of the global information and communication explosion is driven by more, and how much by better technology? *Journal of the Association for Information Science and Technology*, 65(4), 856–861. <https://doi.org/10.1002/asi.23031>.
- Hilbert, M. (2014b). Technological information inequality as an incessantly moving target: The redistribution of information and communication capacities

between 1986 and 2010. *Journal of the Association for Information Science and Technology*, 65(4), 821–835. <https://doi.org/10.1002/asi.23020>.

- Hilbert, M. (2015). *Quantifying the data deluge and the data drought* (SSRN scholarly paper no. ID 2984851). Rochester: Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=2984851>.
- Hilbert, M. (2016). The bad news is that the digital access divide is here to stay: Domestically installed bandwidths among 172 countries for 1986–2014. *Telecommunications Policy*, 40(6), 567–581. <https://doi.org/10.1016/j.telpol.2016.01.006>.
- Hilbert, M., & López, P. (2011). The world's technological capacity to store, compute, and compute information. *Science*, 332(6025), 60–65. <https://doi.org/10.1126/science.1200970>.
- Hilbert, M., & López, P. (2012a). How to measure the world's technological capacity to communicate, store and compute information? Part I: Results and scope. *International Journal of Communication*, 6, 956–979.
- Hilbert, M., & López, P. (2012b). How to measure the world's technological capacity to communicate, store and compute information? Part II: Measurement unit and conclusions. *International Journal of Communication*, 6, 936–955.
- ITU (International Telecommunication Union). (2015). *World Telecommunication/ICT Indicators Database*. Geneva: International Telecommunication Union. Retrieved from <http://www.itu.int/ITU-D/ict/statistics/>.
- López, P., & Hilbert, M. (2012). *Methodological and statistical background on the world's technological capacity to store, communicate, and compute information* (online document). Retrieved from <http://www.martinhilbert.net/WorldInfoCapacity.html>.
- Ookla. (2014). *NetIndex source data*. Retrieved from <http://www.netindex.com/source-data/>.

Communications

Alison N. Novak

Department of Public Relations and Advertising,
Rowan University, Glassboro, NJ, USA

There is much debate about the origins and history of the field of Communications. While many researchers point to a rhetorical origin in ancient Greece, others suggest the field is much newer, developing from psychology and propaganda studies of the 1940s. The discipline includes scholars exploring subtopics such as political communication, media effects, and organizational relationships. The field generally uses both qualitative and quantitative approaches, as well as developing a variety of mixed-methods techniques to understand social phenomena.

Russell W. Burns argues that the field of Communications developed from a need to explore the ways in which media influenced people to behave, support, or believe in a certain idea. Much of Communication studies investigates the idea of media and texts, such as newspaper discourses, social media messages, or radio transcripts. As the field has developed, it has investigated new technologies and media, including those still in their infancies.

Malcom R. Parks states that the field of Communications has not adopted one set definition of big data, but rather sees the term as a means to identify datasets and archival techniques. Singularly thinking of big data as a unit of measurement or a size fails to underscore the many uses and methods used by Communications to explore big datasets.

One frequent source of big data analysis in Communications is that of network analysis or social network analysis. This method is used to explore the ways in which individuals are connected in physical and digital spaces. Communications research on social networks particularly investigates how close individuals are to each other, whom they are connected through, and what resources can be shared amongst networks.

These networks can be archived from social networking sites such as Twitter or Facebook, or alternatively can be constructed through surveys of people within a group, organization, or community. The automated data aggregation of digital social networks makes the method appealing to Communications researchers because it produces large networks quickly and with limited possibility of human error in recording nodes. Additionally, the subfield of Health Communications has adopted the integration of big datasets in an effort to study how healthcare messages are spread across a network.

Natural language processing is another area of big data inquiry in the field of Communications. In this vein of research, scholars explore the way that computers can develop an understanding of language and generate responses. Often studied along with Information Science researchers and Artificial intelligence developers, natural language processing draws from Communications association with linguistics and modern languages. Natural language processing is an attempt to build communication into computers so they can understand and provide more sender-tailored messages to users.

The field of communication has also been outspoken about the promises levied with big data analytics as well as the ethics of big data use. Recognizing that the field is still early in its development, scholars point to the lifespan of other technologies and innovations as examples of how optimism early in the lifecycle often turns into critique. Pierre Levy is one Communications scholar who explains that although new datasets and technologies are viewed as positive changes with big promises early in their trajectory, as more information is learned about their effects, scholars often begin to challenge their use and ability to provide insight.

Communications scholars often refer to big data as the “datafication” of society, meaning turning everyday interactions and experiences into quantifiable data that can be segmented and analyzed using brad techniques. This in particular refers to analyzing data that has not been previously viewed as data before. Although this is partially where the value of big data develops from, for

Communications researchers, this complicates the ability to think holistically or qualitatively.

Specifically, big datasets in Communications research include information taken from social media sites, health records, media texts, political polls, and brokered language transcriptions. The wide variety of types of datasets reflects the truly broad nature of the discipline and its subfields.

Malcom Parks offers suggestions on the future of big data research within the field of Communications. First, the field must situate big data research with larger theoretical contexts. One critique of the data-revolution is the false identification of this form of analysis as being new. Rather than consider big data as an entirely new phenomena, by situating it within a larger history of Communications theory, more direct comparisons between past and present datasets can be drawn.

Second, the field requires more attention to the topic of validity in big data analysis. While quantitative and statistical measurements can support the reliability of a study, validity asks researchers to provide examples or other forms of support for their conclusions. This greatly challenges the ethical notions of anonymity in big data, as well as the consent process for individual protections. This is one avenue in which the quality of big data research needs more work within the field of communications.

Communications asserts that big data is an important technological and methodological advancement within research, however, due to its newness, researchers need to exercise caution when considering its future. Specifically, researchers must focus on the ethics of inclusion in big datasets, along with the quality of analysis and long term effects of this type of dataset on society.

Further Reading

- Burns, R. W. (2003). *Communications: An international history of the formative years*. New York: IEE History of Technology Series.
- Levy, P. (1997). *Collective intelligence: Mankind's emerging world in cyberspace*. New York: Perseus Books.
- Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication*, 64, 355–360.

Community Management

- [Content Moderation](#)
-

Community Moderation

- [Content Moderation](#)
-

Complex Event Processing (CEP)

Sandra Geisler

Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

Synonyms

[Complex event recognition](#); [Event stream processing](#)

Overview

In the CEP paradigm simple and complex events can be distinguished. A *simple event* is the representation of a real-world occurrence, such as a sensor reading, a log entry, or a tweet. A *complex event* is a composite event or also called *situation* which has been detected by identifying a pattern based on the input stream values which may constitute either simple or complex events. As an example for a situation, we consider detecting a material defect in a production line process based on the specific reading values of multiple hardware sensors, that is, the thickness and the flexibility. If the thickness is less than 0.3 and at the same time flexibility is higher than 0.8, a material defect has been detected. An event is characterized by a set of attributes and additionally contains one or more timestamps indicating the time the event has been produced (either assigned by the original source or by the CEP system) or the

duration of its validity, respectively. In our example the complex event could look like this: defect (timestamp,partid). The timestamps play a very important role as especially in CEP systems their time-based relationship in terms of order or parallel occurrence may be crucial to detect a certain complex event. An event is instantiated when the attributes are filled with concrete values and can be represented in different data formats, for example, a relational schema or a nested schema format. This instantiation is also called a *tuple*. Events with the same set of attributes are subsumed under an *event type*, for example, defect. Events are created by an *event producer* which observes a source, for example, a sensor. A potentially unbounded series of events coming from the same source is termed as event stream, which may contain events of different types (*heterogeneous event stream*) or contains only the same event type (*homogeneous event stream*) (Etzion and Niblett 2011). The events are pipelined into an event network consisting of operators processing the events and routing them according to the task they fulfill. These operators usually comprise tasks to filter and transform events and detect patterns on them. The network represents *rules* which determine which complex events should be detected and what should happen if they have been detected. The rule processing can be separated into two phases: detection phase (does the event lead to a rule match?) and production phase (what is the outcome, what has to be done if a rule matched?) (Cugola and Margara 2012b). Usually, also a history of events is kept to represent the context of an event and to keep partial results (if a rule has not been completely matched yet). Finally, *event consumers or sinks* wait for notifications that a complex event has been detected, which in our example could be a monitoring system alerting the production manager or it may be used to sum up the number of defects in a certain time period. Hence, CEP systems are also regarded as extensions of the Publish-Subscribe scheme (Cugola and Margara 2012b), where producers publish data and consumers simply filter out the data relevant to them. In some systems, the produced complex events can serve as an input to further rules enabling *event*

hierarchies and *recursive processing*. CEP is also related to the field of Data Stream Management Systems (DSMS). DSMS constitute a more general way to process data streams and enable generic user queries, while CEP systems focus on the detection of specific contextual composite occurrences using ordering relationships and patterns.

Key Research Findings

Query Components

As described the filtering, transformation, and pattern detection on events can be expressed using operators usually combined in forms of rules. The term *rule* is often interchangeably used with the term query. Depending on the system implementation rules may be added at design time or run time. Rules are either evaluated when an event arrives (event-based) or on a regular temporal basis (time-based).

The filtering of events only allows certain events of interest to participate in the following processing steps. The filter operator is either applied as part of another operator or before the next operators (as a separate operator). A filter expression can be applied to the metadata of an event or the content of an event and can be stateless or stateful. Stateless filters only process one event at a time. Stateful filters are applied to a certain time context (i.e., a time window) where, for example, the first x elements, the most recent x elements, or a random set of elements in the time context, are dropped to reduce the amount of data flowing through the system. This is often done in conjunction with performance measures and policies to fulfill latency requirements (also called *load shedding*).

Transformation operators take an input and produce different result events based on this input. They can also either be stateless (only one input at a time is processed) or stateful (based on multiple input elements). Common operators are projection, value assignment to attributes, copying, modifying, or inserting new attributes, data enrichment, splitting into several event streams, and merging of event streams (join).

To define pattern detection mechanisms, a set of common operators is used. Usually, conjunction (all events happen in the specified time), disjunction (at least one event happens in the specified time), sequence (the events happen sequentially), Kleene operator (the event may happen zero or multiple times), or negation (the event does not occur in the specified time) are used to formulate queries to detect composite events in a specific time frame. So-called *functor patterns* apply aggregation functions such as average, min, max, count, standard deviation, etc., and compare the results against a certain threshold (Etzion and Niblett 2011). Other patterns select events based on values of an attribute in a top k manner in the set of events to be processed in this step. Finally, patterns with regard to time and space, so-called *dimensional patterns*, can be defined (Etzion and Niblett 2011). Temporal patterns include the detection of a sequence of events, the top k events in terms of time (the most recent k or first k events in a time period), trend patterns (a set ordered by time fulfills a criterion, e.g., a value is increasing, decreasing, remains the same, etc.). Spatial patterns are applied to spatial attributes and may fire when events fulfill a certain spatial pattern, such as a minimum, maximum, or average distance between two events, and can also be combined with temporal aspects, such as detecting spatial trends over time.

An often required feature of a query language is also the combination of streaming data with static, historical data. Some query languages offer different representations for such inputs, such as CQL or StreamSQL. The definition of selection policies, that is, if a rule is fired only once, k times, or each time when a pattern has been matched (Cugola et al. 2015), may also be of interest to control the production of events. Similarly, some languages support to restrict events as input, if they do not fulfill a certain contiguity, for example, the rule is only matched, if two combined events are contiguous (Flouris et al. 2017). Additionally, consumption policies can be defined, which control if an input event may be used in the same rule pattern again or if it should be forgotten after it has led to a match.

Query Languages

There are different ways to distinguish the available CEP query languages and how rules are expressed. On the one hand, the way how complex events are retrieved, that is, declaratively or imperatively, can be distinguished. Declarative languages describe what should be the result of the query and these are often variants of SQL or very similar to it, such as the Continuous Query Language (CQL) or the SASE+ language. Imperative languages describe how a result is retrieved and are often manifested in code or visual representations of it, that is, operators are visual components which can be connected to each other to process the event streams. On the other hand, languages can be distinguished based on their theoretical foundation to describe the most important operators. Etzion and Niblett differentiates roughly stream-oriented (e.g., CQL) and rule-oriented languages (e.g., Drools).

Eckert, Bry, Brodt, Poppe, and Hausmann (2011) define more fine-granulated language categories which we will summarize here briefly. Composition-based languages use a combination or nesting of operators to compose events, such as conjunction, negation, and disjunction, to detect events in a certain time frame. A well-known example for this category is the SASE+ language (<http://avid.cs.umass.edu/sase>). Data stream management languages are mainly SQL-based declarative languages which can be used to define CEP style queries in DSMS. The Continuous Query Language (CQL) (Arasu et al. 2006) is a famous deputy of this class comprising the aforementioned operators. It is used in various systems, such as the STREAM system, Oracle CEP, or an extension of Apache Spark (<https://github.com/Samsung/spark-cep>). Other languages of this type are ESL, Esper, StreamInsight (LINQ), or SPADE. Further language types contain state-machine based languages which utilize formalizations of finite state machines to describe rules detecting the pattern. The events may lead to state transitions (in a certain order) where the complex event is detected when a specific state is reached. Production rule-based languages define the pattern in terms of if-then-rules (usually using a higher level programming language such

as Java) where events are represented as facts or objects and are matched against defined rules. An example is the Business Rules Management System Drools (<https://www.drools.org>). If a match is found a corresponding action is evoked (namely creation of the complex event). Finally, languages based on logical languages, such as Prolog, allow for the definition of queries and pattern recognition tasks using corresponding rules and facts.

Time and Order

We already emphasized the prominent status of timestamps as time and may play an important role in detecting complex events expressed in the variety of temporal and spatiotemporal operators. A timestamp is always handled as a specific attribute which is not part of the common attribute set, for example, in the TESLA language (Cugola et al. 2015). A monotonic domain for time can be defined as an ordered, infinite set of discrete time instants. For each timestamp exists a finite number of tuples (but it can also be zero). In literature, there exist several ways to distinguish where, when, and how timestamps are assigned. First of all, the temporal domain from which the timestamps are drawn can be either a logical time domain or a physical clock-time domain. Logical timestamps can be simple consecutive integers, which do not contain any date or time information, but just serve for ordering. In contrast, physical clock-time includes time information (e.g., using UNIX timestamps). Furthermore, systems differ in which timestamps they accept and use for internal processing (ordering and windowing). In most of the systems implicit timestamps, also called internal timestamps or system timestamps, are supported. Implicit timestamps are assigned to a tuple, when it arrives at the CEP system. This guarantees that tuples are already ordered by arrival time, when they are pipelined through the system. Implicit timestamps also allow for estimating the timeliness of the tuple when it is output. Besides a global implicit timestamp (assigned on arrival), there exists also the concept of new (local) timestamps assigned at the input or output of each operator (time of tuple creation). In contrast, explicit timestamps, external timestamps, or application timestamps are created by the sources

and an attribute of the stream schema is determined to be the timestamp attribute. Additionally, in the Stream Mill language ESL there exists the concept of latent timestamps. Latent timestamps are assigned on demand (lazily), that is, only for operations dependent on a timestamp such as windowed aggregates, while explicit timestamps are assigned to every tuple. An interesting question is how timestamps should be assigned to results of for example binary operators and aggregates to ensure semantic correctness. The first option is to use the creation time of an output tuple when using an implicit timestamp model. The second option is to use the timestamp of the first stream involved in the operator, which is suited for explicit and implicit timestamp models. For aggregates similar considerations can be made. For example, if a continuous or windowed minimum or maximum is calculated, the timestamp of the maximal or minimal tuple, respectively, could be used. When a continuous sum or count is calculated, the creation time of the result event or the timestamp of the latest element included in the result can be used. If an aggregate is windowed, there exist additional possibilities. The smallest or the highest timestamp of the events in the window can be used as they reflect the oldest timestamp or most recent timestamp in the window, respectively. Both maybe interesting, when timeliness for an output tuple is calculated, but which one to use depends obviously on the desired outcome. Another possibility would be to take the median timestamp of the window.

Many of the systems and their operators rely on (and assume) the ordered arrival of tuples in increasing timestamp order to be semantically correct (also coined as the *ordering requirement*). But as already pointed out, this cannot be guaranteed especially for explicit timestamps and data from multiple sources. In the various systems basically two main approaches to the problem of disorder have been proposed. One approach is to tolerate disorder in controlled bounds. For example, a slack parameter is defined for order-sensitive operators denoting, how many out-of-order tuples may arrive between the last and the next in-order event. All further out-of-order tuples will be discarded. The second way

to handle disorder is to dictate the order of tuples and reorder them if necessary. While the use of implicit timestamps is a simple way of ordering tuples on arrival, the application semantics often requires the use of explicit timestamps though. Heartbeats are events sent with the stream including at least a timestamp. These markers indicate to the processing operators that all following events have to have a timestamp greater than the timestamp in the punctuation. Some systems buffer elements and output them in ascending order as soon as a heartbeat is received, that is, they are locally sorted. The sorting can be either integrated in a system component or be a separate query operator. Heartbeats are only one possible form of punctuation. Punctuations, in general, can contain arbitrary patterns which have to be evaluated by operators to true or false. Therefore, punctuations can also be used for approximation. They can limit the evaluation time or the number of tuples which are processed by an otherwise blocking or stateful operator. Other methods are, for example, compensation-based techniques, where operators in the query are executed in the same way as if all events would be ordered or approximation-based techniques, where either streams are summarized and events are approximated or approximation is done on a recent history of events (Giatrakos et al. 2019).

Rule Evaluation Strategies

There are basically two strategies to evaluate rules defined in the CEP system (Cugola et al. 2015). Either the rules are evaluated incrementally on each incoming event or the processing is delayed until events in the history fulfill all conditions. The latter requires that all primitive events are stored until a rule fires which may reduce latency (Cugola and Margara 2012a). The first option is the usual case and there have been proposed different strategies how the partial matches are stored (Flouris et al. 2017). Many systems use non-deterministic finite automata or finite state machines, where each state represents a partial or complete match and incoming events trigger state transitions. Further structures comprise graphs, where the leafs represent the simple events which are incrementally combined to

partial matches representing the inner nodes. The root node constitutes the overall match of the rule. Similarly, events may be pipelined through operator networks which forward and transform the events based on their attributes (Etzion and Niblett 2011). Finally, graphs are also used as a combination of all activated rules to detect event dependencies (Flouris et al. 2017).

Further Directions for Research

Uncertainty in CEP

Uncertainty has been studied intensively for Database Management Systems and several systems implementing probabilistic query processing have been proposed. Also for CEP this is an interesting aspect as it often handles data from erroneous sources and uncertainty may be considered on various levels. For data streams in general Kanagal and Deshpande distinguished two main types of uncertainties. First, the existence of a tuple in a stream can be uncertain (how probable is it for this tuple to be present at the current time instant?), which is termed *tuple existence uncertainty*. Second, the value of an attribute in a tuple can be uncertain, which is called *attribute value uncertainty* (what is the probability of attribute X to have a certain value?). The latter aspect is crucial as many data sources may inherently create erroneous data, such as sensors or other devices or are estimates of a certain kind. Naturally, both aspects have also been considered specifically for CEPs (Flouris et al. 2017; Cugola et al. 2015). There are two ways to represent attribute value uncertainty. The attribute value can be modeled as a random variable which is accompanied by a corresponding probability density function (pdf) describing the deviation from the exact value. The second option is to attach to each attribute value a concrete probability value. In consequence, the tuple existence uncertainty to consist of a certain configuration of values may be modeled by a joint distribution function, which multiplies the probabilities of the single attribute values. Finally, each event may then have a value which indicates the probability of its occurrence (Cugola et al. 2015). Depending on the

assumption if attributes are independent of each other, probability values can be propagated to complex events depending on the operators applied (combination, aggregation etc.). A further instance of this problem is temporal uncertainty. The time of the occurrence of an event, synchronization problems between clocks of different processing nodes, or different granularities of event occurrences may be observed. Zhang, Diao, and Immerman introduced a specific temporal uncertainty model to express the uncertain time of occurrence by an interval. Another level of uncertainty can be introduced in the rules. For example, Cugola et al. use Bayesian networks to model the relationships between simple and complex events in the rules to reflect the uncertainty which occur in complex systems. How uncertainty is handled on data and rule level can be again categorized based on the theoretical foundation. Alevizos, Skarlatidis, Artikis, and Paliouras (2017) distinguish between automata-based approaches, first-order logic and graph models, petri nets, and syntactical approaches using grammars.

Rule Learning

An interesting direction to follow in CEP is the automatic learning of rules. On the one hand, the definition of rules is usually a lengthy manual task, which is done by domain experts and data scientists. It is done at design time and may need multiple cycles to adapt the rule to the use case at hand. On the other hand, depending on the data, the rules to detect certain complex events might not be obvious. Hence, it might not be possible to define a rule from scratch or in a reasonable time. In both cases, it is desirable to learn rules from sample data to support the process of rule definition. This can be done by using labeled historical data. Margara et al. (2014) identified several aspects, which need to be considered to learn a rule, for example, the time frame or the event types and attributes to be considered. Consequently, they build custom learners for each identified subproblems (or constraints) and use labeled historical data divided into positive and negative event traces which either match or do not match a complex event. Other approaches use machine

learning approaches specifically suited for data streams. For example, Mehdiyev et al. (2015) compare different rule-based classifiers to detect event patterns for CEP to derive rules for activity detection based on accelerometer data in phones. Usually, CEP is reactive, that is, the detected complex events lie in the past. Mousheimish, Taher, and Zeitouni present an approach how predictive rules for CEP could be learned, such that events in the near future could be predicted. They use mining on labeled multivariate time series to create rules which are installed online, that is, are activated during run time.

Scalability

As CEPs are systems which operate on data streams adaptability to varying workloads is important. Looking at the usual single centralized systems, there are a multiple levels which can be considered. If the input load is too high and completeness of input data may not be of major importance, data may be sampled to decrease the system load. This can be done by measuring system performance using QoS parameters, such as the output latency or the throughput. Based on these measures, for example, load shedders integrated in the event processor or directly into operator implementations may drop events when performance cannot be kept on an acceptable level. Besides load balancing and load shedding, parallelization techniques can be applied to increase the performance of a system. Giatrakos et al. (2019) distinguish two kinds of parallelization for CEP, namely, task and data parallelization, where task parallelization comprises the distribution of queries and subqueries or single operators to different nodes where they are executed. Data parallelization considers the distribution of data to multiple instances of the same operator or query. A CEP system should be scalable in terms of queries as some application contexts can get complex and require the introduction of several queries at the same time. Hence, for parallelization the execution of multiple queries can be distributed to different processing units. Furthermore, a query can be divided into subqueries or single operators and the parts can be distributed over multiple threads parallelizing their

execution, bringing up the need for intra- and multiquery optimization, for example, by sharing results of operators in an overall query plan. In data parallelization the data is partitioned and distributed to equal instances of operators or subqueries and the results are merged in the end. A possible strategy to implement CEPs in a scalable way is elevating CEP systems to big data platforms as these are designed to serve high workloads by workload distribution and elastic services. Giatrakos et al. (2019) show how CEP can be integrated with Spark Streaming, Apache Flink, and Apache Storm taking advantage of the corresponding abilities for scalability.

Cross-References

- [Big Data Quality](#)
- [Data Processing](#)
- [Data Scientist](#)
- [Machine Learning](#)
- [Metadata](#)

Further Reading

- Alevizos, E., Skarlatidis, A., Artikis, A., & Paliouras, G. (2017). Probabilistic complex event recognition: A survey. *ACM Computing Surveys (CSUR)*, 50(5), 71.
- Arasu, A., Babu, S., & Widom, J. (2006). The CQL continuous query language: Semantic foundations and query execution. *The VLDB Journal*, 15(2), 121–142.
- Cugola, G., & Margara, A. (2012a). Low latency complex event processing on parallel hardware. *Journal of Parallel and Distributed Computing*, 72(2), 205–218.
- Cugola, G., & Margara, A. (2012b). Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR)*, 44(3), 15.
- Cugola, G., Margara, A., Matteucci, M., & Tamburrelli, G. (2015). Introducing uncertainty in complex event processing: Model, implementation, and validation. *Computing*, 97(2), 103–144. <https://doi.org/10.1007/s00607-014-0404-y>.
- Eckert, M., Bry, F., Brodt, S., Poppe, O., & Hausmann, S. (2011). A cep babelfish: Languages for complex event processing and querying surveyed. In *Reasoning in event-based distributed systems* (pp. 47–70). Springer, Berlin, Heidelberg.
- Etzion, O., & Niblett, P. (2011). *Event processing in action*. Greenwich: Manning.
- Flouris, I., Giatrakos, N., Deligiannakis, A., Garofalakis, M., Kamp, M., & Mock, M. (2017). Issues in complex event processing: Status and prospects in the big data era. *Journal of Systems and Software*, 127, 217–236.
- Giatrakos, N., Alevizos, E., Artikis, A., Deligiannakis, A., & Garofalakis, M. (2019). Complex event recognition in the big data era: A survey. *The VLDB Journal*, 29, 313. <https://doi.org/10.1007/s00778-019-00557-w>.
- Kanagal, B., & Deshpande, A. (2009). Efficient query evaluation over temporally correlated probabilistic streams. In *IEEE 25th international conference on data engineering, 2009. icde'09* (pp. 1315–1318).
- Luckham, D. C., & Frasca, B. (1998). *Complex event processing in distributed systems* (Technical report) (Vol. 28). Stanford: Computer Systems Laboratory, Stanford University.
- Margara, A., Cugola, G., & Tamburrelli, G. (2014). Learning from the past: Automated rule generation for complex event processing. In *Proceedings of the 8th ACM international conference on distributed event-based systems* (pp. 47–58).
- Mehdiyev, N., Krumeich, J., Enke, D., Werth, D., & Loos, P. (2015). Determination of rule patterns in complex event processing using machine learning techniques. *Procedia Computer Science*, 61, 395–401.
- Mousheimish, R., Taher, Y., & Zeitouni, K. (2017). Automatic learning of predictive cep rules: Bridging the gap between data mining and complex event processing. In *Proceedings of the 11th ACM international conference on distributed and event-based systems* (pp. 158–169).
- Zhang, H., Diao, Y., & Immerman, N. (2010). Recognizing patterns in streams with imprecise timestamps. *Proceedings of the VLDB Endowment*, 3(1–2), 244–255.

Complex Event Recognition

- [Complex Event Processing \(CEP\)](#)

Complex Networks

Ines Amaral

University of Minho, Braga, Minho, Portugal
 Instituto Superior Miguel Torga, Coimbra,
 Portugal

Autonomous University of Lisbon, Lisbon,
 Portugal

In recent years, the emergence of a large amount of data dispersed in several types of databases enabled the extraction of information on a never seen scale. Complex networks allow the

connection of a vast amount of scattered and unstructured data in order to understand relations, construct models for their interpretation, analyze structures, detect patterns, and predict behaviors.

The study of complex networks is multidisciplinary and covers several knowledge areas as computer science, physics, mathematics, sociology, and biology. Within the context of the Theory of Complex Networks, a network is a graph that represents a set of nodes connected by edges, which together form a network. This network or graph can represent relationships between objects/agents. Graphs can be used to model many types of relations and processes in physical, biological, social, and information systems.

A graph is a graphical representation of a pattern of relationships and is used to reveal and quantify important structural properties. In fact, the graphs identify structural patterns that cannot be detected otherwise. The representation of a network or a graph consists of a set of nodes (vertices) that are connected by lines, which may be arcs or edges, depending on the type of relationship to study. Matrices are an alternative to represent and summarize data networks, containing exactly the same information as a graph.

Studies of complex networks have their origin in Graph Theory and in Network Theory. In the eighteenth century, the Swiss mathematician Euler developed the founding bases of Graph Theory. Euler solved the problem of the bridges of Königsberg through the modeling of a graph that transformed the ways in straight lines and their intersection points. It is considered that this was the first graph developed. The primacy of relations is explained in the work of George Simmel, a German sociologist who is often nominated as the theoretical antecedent of Network Analysis. Simmel argued that the social world was the result of interactions and not the aggregation of individuals. The author argued that society was no more than a network of relationships, given the intersection of these as the basis for defining the characteristics of social structures and individual units.

The modeling of complex networks is supported by the mathematical formalism of

Graph Theory. Studying the topology of networks through the Graph Theory, formalists' authors pursue to analyze situations in which the phenomena in question establish relations among themselves. The premise is that everything is connected and nothing happens in isolation, which is based on the formalist perspective of Network Theory.

Network Theory approaches the study of graphs as a representation of either symmetric or asymmetric relations relationships between objects. This theory assumes the perspective that social life is relational, which suggests that the attributes by itself have no meaning that can explain social structures or other networks. The focus of the analysis is the relationships established in a given system. Thus, the purpose of different methodologies within the Network Theory is to detect accurately and systematically patterns of interaction.

In the formalist perspective of Network, a system is complex when their properties are not a natural consequence of their isolated elements. In this sense, the theoretical propose is the application of models in order to identify common patterns of interaction systems. The network models designed by the authors of formalist inspiration have been used in numerous investigations and can be summarize three different perspectives: random networks, small-world networks and scale-free networks.

The model of random networks was proposed by Paul Erdős and Alfred Rényi in 1959 and is considered the simplest model of complex systems. The authors argued that the process of formation of networks was random. Assuming as true the premise that nodes aggregate randomly, the researchers concluded that all actors in a network have a number of close links and the same probability of establish new connections. The theory focuses on the argument that the more complex is the network, the greater is the probability of their construction is random. From the perspective of Erdős and Rényi, the formation of networks is based on two principles: equality or democracy of networks (all nodes have the same probability of belonging to the network) and the transition (from isolation to connectivity).

Barabási (2003) and other authors argue that the theory of randomness cannot explain the complex networks that exist in the world today.

Watts and Strogatz proposed the small-world model in 1998. The model assumes as its theoretical basis the studies of “small worlds” of Milgram, who argued that 5.2 degrees of separation mediate the distance between any two people in the world, and Granovetter’s theories on the weak social ties between individuals and the structural importance, and the influence they have on the evolution and dynamics of networks. The researchers created a model where some connections were established by proximity and others randomly, which transforms networks into small worlds. Watts and Strogatz found that the separation increases more slowly than the network evolution. This theory, called effect of “small world” or “neighborhood effect,” argues that in contexts where there are very close connected members, actors bind so that there are few intermediaries. Therefore, there is a high degree of clustering and a reduced distance between the nodes. According to the model developed by Watts and Strogatz, the average distance between any two people would not exceed a small number of other people, being only required that there were a few random links between groups.

In a study that sought to assess the feasibility of applying the Theory of Small Worlds to the World Wide Web, Barabási and Albert demonstrated that networks are not formed randomly. The researchers have proposed the model of “scale-free networks,” which is grounded on the argument that networks that evolve are based on mechanisms of preferential attachment. As Granovetter’s theories and studies of Watts and Strogatz, Barabási and Albert argued that there is an order in the dynamic structure of networks and defined the preferential attachment as a standard for pattern of structuring such as “rich get richer”. Therefore, many more connections a node has, the higher is the probability of having more links. The model of scale-free networks is based on growth and preferential attachment. In this type of network, the main feature is the unequal distribution

of connections among agents, and the trend for the new nodes being connected to others who have high degree of connectivity. Power laws are associated to this specific symmetry.

The Theory of Complex Networks has emerged in the 1990s of the last century due to the Internet and computers capable of processing big data. Despite the similarities, the Theory of Complex Networks differs from Graph Theory, in three basic aspects: (i) it is related to the modeling of real networks, through analysis of empirical data; (ii) the networks are not static, but evolve over time, changing its structure; (iii) the networks are structures where dynamic processes (such as the spread of viruses or opinions) can be simulated.

The Theory of Complex Networks is currently widely applied both to the characterization and the mathematical modeling of complex systems. Complex networks can be classified according to statistical properties as the degree or the coefficient of aggregation. There are several tools to generate graphical representations of networks. Visualization of complex networks can represent large-scale data and enhance their interpretation and analysis.

Big data and complex networks share three proprieties: large scale (volume), complexity (variety), and dynamics (velocity). Big data can change the definition of knowledge, but by itself is not self-explanatory. Therefore, the ability to understand, model, and predict behavior using big data can be provided by the Theory of Complex Networks.

As mathematical models of simpler networks do not display the significant topological features, modeling big data in complex networks can facilitate the analysis of multidimensional networks extracted from massive data sets. The clustering of data in networks provides a way to understand and obtain relevant information from large data sets, which allows learning, inferring, predicting, and having knowledge of large volumes of dynamic data sets.

Complex networks may promote collaboration between many disciplines towards large-scale

information management. Therefore, computational, mathematical, statistical, and algorithmic techniques can be used to modeling high dimensional data, large graphs, and complex data in order to detect structures, communities, patterns, locations, influence, and model transmissions in interdisciplinary research at the interface between big data analysis and complex networks. Several areas of knowledge can benefit from the use of complex networks models and techniques for analyzing big data.

Cross-References

- [Computational Social Sciences](#)
- [Data Visualization](#)
- [Graph-Theoretic Computations/Graph Databases](#)
- [Network Analytics](#)
- [Network Data](#)
- [Social Network Analysis](#)
- [Visualization](#)

Further Reading

- Barabási, A.-L. (2003). *Linked*. Cambridge, MA: Perseus Publishing.
- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509.
- Bentley, R. A., O'Brien, M. J., & Brock, W. A. (2014). Mapping collective behavior in the big-data era. *Behavioral and Brain Sciences*, 37, 63.
- Boccaletti, S., et al. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4–5), 175.
- McKelvey, K., et al. (2012). Visualizing communication on social media: Making big data accessible. arXiv preprint arXiv:1202.1367.
- Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825), 268.
- Watts, D. (2003). *Six degrees: The science of a connected age*. New York: Norton.
- Watts, D. (2004). The “new” science of networks. *Annual Review of Sociology*, 30(1), 243.

Computational Social Sciences

Ines Amaral

University of Minho, Braga, Minho, Portugal
Instituto Superior Miguel Torga, Coimbra,
Portugal

Autonomous University of Lisbon, Lisbon,
Portugal

Computational social sciences is a research discipline at the interface between computer science and the traditional social sciences. This interdisciplinary and emerging scientific field uses computationally methods to analyze and model social phenomena, social structures, and collective behavior. The main computational approaches to the social sciences are social network analysis, automated information extraction systems, social geographic information systems, complexity modeling, and social simulation models.

New areas of social science research have arisen due the existence of computational and statistical tools, which allow social scientists to extract and analyze large datasets of social information. Computational social sciences diverges from conventional social science because of the use of mathematical methods to model social phenomena. As an intersection of computer science, statistics, and the social sciences, computational social science is an interdisciplinary subject, which uses large-scale demographic, behavioral, and network data to analyze individual activity, collective behaviors, and relationships. Modern distributed computing frameworks, algorithms, statistics, and machine learning methods can improve several social science fields like anthropology, sociology, economics, psychology, political science, media studies, and marketing. Therefore, computational social sciences is an interdisciplinary scientific area, which explores social dynamics of society through advanced computational systems.

Computational social science is a relatively new field, and its development is closely related

Computational Ontology

- [Ontologies](#)

to the computational sociology that is often associated to the study of social complexity, which is a useful conceptual framework for the analysis of society. Social complexity is theory neutral that frames both local and global approaches to social research. The theoretical background of this conceptual framework dates back to the work of Talcott Parsons on action theory, the integration of the study of social order with the structural features of macro and micro factors. Several decades later, in the early 1990s, social theorist Niklas Luhmann began to work on the themes of complex behavior. By then, new statistical and computational methodologies were being developed for social science problems.

Nigel Gilbert, Klaus G. Troitzsch, and Joshua M. Epstein are the founders of modern computational sociology, merging social science research with simulation techniques in order to model complex policy issues and essential features of human societies. Nigel Gilbert is a pioneer in the use of agent-based models in the social sciences. Klaus G. Troitzsch introduces the method of computer-based simulation in the social sciences. Joshua M. Epstein developed, with Robert Axtell, the first large-scale agent-based computational model, which aims to explore the role of social experiences such as seasonal migrations, pollution, and transmission of disease.

As an instrument-based discipline, computational social sciences enables the observation and empirical study of phenomena through computational methods and quantitative datasets. Quantitative methods such as dynamical systems, artificial intelligence, network theory, social network analysis, data mining, agent-based modeling, computational content analysis, social simulations (macrosimulation and microsimulation), and statistical mechanics are often combined to study complex social systems.

Technological developments are constantly changing society, ways of communication, behavioral patterns, the principles of social influence, and the formation and organization of groups and communities, enabling the emergence of self-organized movements. As technology-mediated behaviors and collectives are primary elements in the dynamics and in the design of social structures, computational approaches are critical to understand the

complex mechanisms that form part of many social phenomena in contemporary society. Big data can be used to understand many complex phenomena as it offers new opportunities to work toward a quantitative understanding of our complex social systems. Technological-mediated social phenomena emerging over multiple scales are available in complex datasets. Twitter, Facebook, Google, and Wikipedia showed that it is possible to relate, compare, and predict opinions, attitudes, social influences, and collective behaviors. Online and offline big data can provide insights that allow the understanding of social phenomena like diffusion of information, polarization in politics, formation of groups, and evolution of networks.

Big data is dynamic, heterogeneous, and inter-related. But it is also often noisy and unreliable. However, even so, big data may be more valuable to social sciences than small samples because the overall statistics obtained from frequent patterns and correlation analysis disclose often hidden patterns and more reliable knowledge. Furthermore, when big data is connected, it forms large networks of heterogeneous information with data redundancy that can be exploited to compensate for the lack of data, to validate trust relationships, to disclose inherent groups, and to discover hidden patterns and models. Several methodologies and applications in the context of modern social science datasets allow scientists to understand and study different social phenomena, from political decisions to the reactions of economic markets to the interactions of individuals and the emergence of self-organized global movements.

Trillions of bytes of data can be captured by instruments or generated by simulation. Through better analysis of these large volumes of data that are becoming available, there is the potential to make further advances in many scientific disciplines and improve the social knowledge and the success of many companies. More than ever, science is now a collaborative activity. Computational systems and techniques created new ways of collecting, crossing and interconnecting data. Analysis of big data are now at the disposal of social sciences, allowing the study of cases in macro- and in microscales in connection to other scientific fields.

Cross-References

- [Computer Science](#)
- [Data Visualization](#)
- [Network Analytics](#)
- [Network Data](#)
- [Social Network Analysis](#)
- [Visualization](#)

Further Reading

- Banks, S., Lempert, R., & Popper, S. (2002). Making computational social science effective epistemology, methodology, and technology. *Social Science Computer Review*, 20(4), 377–388.
- Bainbridge, W. S. (2007). Computational sociology. In *The Blackwell Encyclopedia of Sociology*. Malden, MA: Blackwell Publishing.
- Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3), 259–271.
- Conte, R., et al. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1), 325–346.
- Lazer, D., et al. (2009). Computational social science. *Science*, 323(5915), 721–723.
- Miller, J. H., & Page, S. E. (2009). *Complex adaptive systems: An introduction to computational models of social life*. Princeton: Princeton University Press.
- Oboler, A., et al. (2012). The danger of big data: Social media as computational social science. *First Monday* 17 (7). Retrieved from <http://firstmonday.org/article/view/3993/3269/>.

Computer Science

Ramón Reichert

Department for Theatre, Film and Media Studies,
Vienna University, Vienna, Austria

Computer science is the scientific approach to the automatic processing of data and information using digital computing machines. Computer science has developed itself at the interface between two scientific disciplines: on the one hand it emerges from the formal logical method of mathematics; on the other hand, it takes genuine problems of engineering sciences and tries to develop

machine-based designs for application-oriented questions.

Computer Science and Big Data

The buzzword “big data” is on everyone’s lips and not only describes scientific data practices but also stands for societal change and a media culture in transition. On the assumption that digital media and technologies do not merely convey neutral messages but establish cultural memory and develop social potency, they may be understood as discourses of societal self-reflection. Over the past few years, big data research has become highly diversified and has yielded a number of published studies, which employ a form of computer-based social media analysis supported by machine-based processes such as text analysis (quantitative linguistics), sentiment analysis (mood recognition), social network analysis, image analysis, or other processes of a machine-based nature. Given this background, it would be ethically correct to regularly enlighten users of online platforms about the computer-based possibilities, processes, and results associated with the collection and analysis of large volumes of data. As a phenomenon and discipline developed only in the past several years, big data can be described as the collection, manipulation, and analysis of massive amounts of data – and the decisions made from that analysis. Moreover, big data is affecting and will affect almost all fields of study, from criminology to philosophy, business, government, transportation, energy, genetics, medicine, physics, and more. As we tackle big data in this proposed encyclopedia, we objectively report on the negative effects of loss of privacy, surveillance, and possible misuse of data in trade-offs for security. On the other hand, it is big data that is helping us to peer into the human genome for invaluable medical insights, or to reach deep across the universe, discovering planets much like our own. In the era of big data, the status of social networks has changed radically. Today, they increasingly act as gigantic data collectors for the observational requirements of social-statistical knowledge and serve as a

prime example of normalizing practices. Where extremely large quantities of data are analyzed, it now usually entails the aggregation of moods and trends. Numerous studies exist where the textual data of social media has been analyzed in order to predict political attitudes, financial and economic trends, psychopathologies, and revolutions and protest movements. The statistical evaluation of big data promises a range of advantages, from increased efficiency in economic management via measurement of demand and potential profit to individualized service offers and better social management.

The structural change generated by digital technologies, as main driver for big data, offers a multitude of applications for sensor technology and biometrics as key technologies. The conquest of mass markets through sensor and biometric recognition processes can sometimes be explained by the fact that mobile, web-based terminals are equipped with a large variety of different sensors. More and more users come this way into contact with the sensor technology or with the measurement of individual body characteristics. Due to the more stable and faster mobile networks, many people are permanently connected to the Internet using their mobile devices, providing connectivity an extra boost. With the development of apps, application software for mobile devices such as smartphones (iPhone, Android, BlackBerry, Windows Phone) and Tablet computer, the application culture of biosurveillance changed significantly, since these apps are strongly influenced by the dynamics of the bottom-up participation. Therefore the algorithmic prognosis of collective processes enjoys particularly high political status, with the social web becoming the most important data-source for knowledge on governance and control.

Within the context of big data, on the other hand, a perceptible shift of all listed parameters has taken place, because the acquisition, modeling, and analysis of large amounts of data, accelerated by servers and by entrepreneurial individuals, is conducted without the users' knowledge or perusal. Consequently, the socially acceptable communication of big data research seeks to integrate the methods, processes, and

models used for data collection in publication strategies, in order to inform the users of online platforms or to invite them to contribute to the design and development of partially open interfaces.

The Emergence of Computer Science

The basic principle of the computer is the conversion of all signs and sign processes in arithmetic operations. In this respect, the history of computer science refers to earlier traditions of thought, which already had an automation of calculations in mind. Before the computer was invented as a tangible machine, there were already operational concepts of the use of symbols, which provided the guidelines for the upcoming development of the computer. The arithmetic, algebraic, and logical calculi of operational use of symbols can be understood as pioneers in computer science. The key-thought behind the idea of formalization is based on the use of written symbols, both schematic and open to interpretation. Inspired by the algebraic methods of mathematics, René Descartes proposed, in his “*Regulae ad directionem ingenii*” in 1628, for the first time the idea of the unity of a rational, reasoned thinking, a *mathesis universalis*. The idea of mathematics as a method of gaining knowledge, that works without object bondage, was taken from the analytic geometry of Pierre de Fermat (1630) and further developed by Gottfried Wilhelm Leibniz in his early work “*Dissertatio de Arte combinatoria*” published in 1666. Leibniz intended his *characteristica universalis* to be a universal language of science and created a binding universal symbolism for all fields of knowledge, which is constructed according to natural sciences and mathematical models. The Boolean algebra of George Boole resulted from the motivation to describe human thinking and action using precise formal methods. With his “*Laws of Thought*” (1854), George Boole laid the foundations of mathematical logic and established in the form of *Boolean algebra*, the fundamental mathematical principles for the whole technical computer science. This outlined development of a secure logical language of

symbols gives form to the operational basis of modern computer technology.

A comparative historical analysis of the data processing, taking into account the material culture of data practices from the nineteenth to the twenty-first century, shows (Gitelman and Pingree 2004) that in the nineteenth century the researcher's interests on the taxonomic knowledge was strongly influenced by the mechanical data practices – long before computer-based methods of data collection even existed (Driscoll 2012). Further studies analyze and compile the social and political conditions and effects of the transition from mechanical data count, of the first census in 1890 through the electronic data processing in the 1950s, to digital social monitoring of the immediate present (Bollier 2010, p. 3). Published in 1937, the work of the British mathematician and logician Alan Mathison Turing *On Computable Numbers, with an Application to the "Entscheidungsproblem,"* in which he developed a mathematical model machine, is to this day of vital importance for the history of theories of modern information and computer technology. With his invention of the *universal Turing machine*, Turing is widely considered to be one of the most influential theorists of the early computer development. In 1946, the mathematician John von Neumann developed the key components of a computer, which is in use until today: control- and arithmetic unit, memory and input/output facilities. During the 1960s the first generation of informatics specialists from the field of social sciences, such as Herbert A. Simon (1916–2001), Charles W. German (1912–1992), and Harold Guetzkow (1915–2008), started to systematically use calculating machines and punch cards for the statistical analysis of their data. (Cioffi-Revilla 2010, pp. 259–271) The computer is as an advanced calculator, which translates the complete information into a binary code and electrically transmits it in form of signals. This way the computer, as a comprehensive hypermedia, is able to store, edit, and deliver, not only verbal texts but also visual and auditory texts in a multimedia convergence space.

The digital large-scale research with its large data processing centers and server farms play,

since the late twentieth century, a central role in the production, processing, and management of computer science knowledge. Concomitantly, media technologies of data collection and processing, as well as media that develop knowledge by using spaces of opportunity move to the center of knowledge production and social control. In this sense we can speak of both *data-based* and *data-driven* sciences, since the production of knowledge has become dependent on the availability of computer technology infrastructures and on the development of digital applications and methods.

Computational Social Science

In the era of big data, the importance of social networking culture has changed radically. Social media acts today as a gigantic data collector and as relevant data sources for the digital communications research: "Social media offers us the opportunity for the first time to both observe human behavior and interaction in real time and on a global scale." (Golder and Macy 2012, p. 7).

The large amounts of data are collected in different domains of knowledge, such as biotechnology, genomics, labor and financial sciences, or trend research, rely in their work and studies on the results of information processing of big data, and formulate on this basis significant models of the current status and future development of social groups and societies. The big data research became significantly differentiated in recent years, as numerous studies have been published, using machine-based methods such as text analysis (quantitative linguistics), sentiment analysis (mood detection), social network analysis, and image analysis or otherwise machine-based processes of computer-based social media analysis.

The newly emerging discipline of "Computational Social Science" (Lazer et al. 2009, pp. 721–723; Conte et al. 2012, pp. 325–346) evaluates the large amounts of data online use in the backend area and has emerged as a new leading science in the study of social media web 2.0. It provides a common platform for computer science and social sciences connecting the different expert opinions

on computer science, society, and cultural processes. The computer science deals with the computer-based elaboration of large databases, which no longer cope with the conventional methods of statistical social sciences. Its goal is to describe the social behavioral patterns of online users on the basis of methods and algorithms of data mining: “To date, research on human interactions has relied mainly on one-time, self-reported data on relationships” (Lazer et al. 2009, p. 722). In order to answer this question of social behavior in a relevant or meaningful way, the computer science requires the methodological input of social sciences. With their knowledge of theories and methods of social activity, social sciences make a valuable contribution to the formulation of relevant issues.

Digital Methods

At the interface between the “computational social science” (Lazer et al. 2009) and the “cultural analytics” (Manovich 2009, pp. 199–212), an interdisciplinary theoretical field has emerged, reflecting the new challenges of the digital Internet research. The representatives of the so-called digital methods pursue the aim of rethinking the research about use (audience research), by interpreting the use practices of the Internet as a cultural change and as social issues (Rogers 2013, p. 63). Analogous methods, though that have been developed for the study of interpersonal or mass communication, cannot simply be transferred to digital communication practices. Digital methods can be understood as approaches that focus on the genuine practice of digital media and not on existing methods adapted for Internet research. According to Rogers (2013), digital methods are research approaches that take advantage of large-scale digital communication data to, subsequently, model and manage this data using computational processes. Both the approach of “Computational Social Science” and the questioning of “Digital Methods” represent the fundamental assumption that by using the supplied data, which creates social

media platforms, new insights into human behavior, into social issues beyond these platforms, and into their software can be achieved. Numerous representatives of computer-based social and cultural sciences sustain the assumption that online data could be interpreted as social *environments*. To do so, they define the practices of Internet use by docking them using a positivist data term, which comprehends the user practices as an expression of specifiable social activity. The social positivism of “Computational Social Science” in social media platforms neglects, however, the meaningful and intervening/instructive role of the media in the production of social roles and stereotyped conducts in dealing with the medium itself. With respect to its postulate of objectivity, social behaviorism of online research can, in this regard, be questioned.

The vision of such native- digital research methodology, whether in the form of a “computational social science” (Lazer et al. 2009, pp. 721–723) or “cultural analytics” (Manovich 2009, pp. 199–212) is, however, still incomplete and requires an epistemic survey of digital methods in Internet research of the following areas:

1. Digital methods as *validity theoretical project*. It stands for a specific process that claims the social recognition of action orientations. The economy of computer science, computational linguistics, and empirical communication sociology not only form a network of scientific fields and disciplines but they also develop, in their strategic collaborative projects, certain expectations, describing and explaining the social world and are, in this respect, intrinsically connected with epistemic and political issues. In this context, the epistemology, questioning the self-understanding of digital methods, deals with the social effectiveness of the digital data science.
2. Digital methods as *constitutional theoretical construct*. The relation to the object in big data research is heterogeneous and consists of different methods. Using interface technologies, the process of data tracking, of keyword tracking, of automatic network analysis, of

argument and sentiment analysis, or machine-based learning results in critical perspectives of data constructs. Against this background, the Critical Code Studies try to make the media techniques, of computer science power relations, visible and study the technical and infrastructural controls over layer models, network protocols, access points, and algorithms.

3. Digital methods may ultimately be regarded as a *founding theoretical fiction*. The relevant research literature has dealt extensively with the reliability and validity of scientific data collection and came to the conclusion that the data interfaces of Social Net (Twitter, Facebook, YouTube) act more or less like dispositive orders according to a gatekeeper. The filter interface generates the API's (*application programming interfaces*) economically motivated exclusionary effects for network research that cannot be controlled by their own efforts.

In this context of problem-oriented development of computer science, the expectations on the science of the twenty-first century have significantly changed. In the debates increasingly claims are being made, they insist in processing historically, socially, and ethically leading aspects of digital data practices – associated with the purpose to anchor these aspects in the future scientific cultures and epistemologies of data generation and data analysis. Lazer et al. demand of future *computer scientists* a responsible use of available data and see in negligent handling a serious threat to the future of the discipline itself: “A single dramatic incident involving a breach of privacy could produce a set of statutes, rules, and prohibitions that could strangle the nascent field of computational social science in its crib. What is necessary, now, is to produce a self-regulatory regime of procedures, technologies, and rules that reduce this risk but preserve most of the research potential.” (Lazer et al. 2009, p. 722) If it is going to be made research on social interaction using computer science and big data, then the responsible handling of data as well as the

compliance with data protection regulations are key issues.

Further Reading

- Bollier, D. (2010). *The promise and peril of big data*. Washington, DC: The Aspen Institute. Online: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.
- Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3), 259–271.
- Conte, R., et al. (2012). Manifesto of computational social science. *European Physical Journal: Special Topics*, 214(1), 325–346.
- Driscoll, K. (2012). From punched cards to ‘Big Data’: A social history of database populism. *Communication + 1*, 1(1), 1. Online: <http://kevindriscoll.info/>.
- Gitelman, L., & Pingree, G. B. (2004). *New media: 1740–1915*. Cambridge, MA: MIT Press.
- Golder, S., & Macy, M. (2012). Social science with social media. *Footnotes*, 40(1), 7. Online: http://www.asanet.org/footnotes/jan12/socialmedia_0112.html.
- Lazer, D., et al. (2009). Life in the network: The coming age of computational social science.” *Science*, 323 (5915), 721–723.
- Manovich, L. (2009). How to follow global digital cultures: Cultural analytics for beginners. In K. Becker & F. Stalder (Eds.), *Deep search: The politics of search beyond Google* (pp. 198–212). Innsbruck, Studienverlag.
- Rogers, R. (2013). *Digital methods*. Cambridge, MA: MIT Press.

Computer-Assisted Reporting

- [Media](#)

Consensus Methods

- [Ensemble Methods](#)

Console

- [Dashboard](#)

Content Management System (CMS)

Yulia A. Strekalova¹ and Mustapha Bouakkaz²

¹College of Journalism and Communications,
University of Florida, Gainesville, FL, USA

²University Amar Telidji Laghouat, Laghouat,
Algeria

The use of content management systems (CMSs) dates back to late 1990s. Taken broadly, CMSs include systems for strategic decision-making in relation to organizational knowledge management and sharing, or, more narrowly, online applications for share this knowledge with internal and external users. CMSs are automated interfaces that allow system users to create, publish, edit, index, and control access to their content without having to learn hypertext markup language (HTML) or other programming languages. CMSs have several possible advantages, like low-cost, built-in pathways for customization and upgrades, flexibility in content access, and ease in use by nontechnical content producers. At the same time, especially in these days of big data and massive datasets, large-scale CMSs can require extended strategic planning and system pre-evaluation. Enterprise-level CMSs may require extensive staff training, hardware investments, and commitments for ongoing maintenance. However, a CMS also may present targets for cyberattacks and security threats when not managed as an integral part of an organization's overall information infrastructure.

Definition

CMS as a concept evolved organically and there are no official standards that guide or prescribe the features CMS should or should not have. Overall, CMS tools aim to manage comprehensive websites and online collaboration portals through the management of a process of collecting,

managing, and publishing content, thus delivering and disseminating knowledge. To that end, a CMS provides a platform for linking and using large datasets in connection to tools for strategic planning and decision-making.

CMS has different meanings and foci across various disciplines and practical applications. The viewpoint of managerial applications of CMS puts more emphasis on the knowledge management and the strategic decisions adopted to manage and deliver business knowledge, while information processing perspective focuses on the process of collecting, managing, and publishing content. More specifically, this latter viewpoint defines a CMS as a computer application that allows to publish, edit, and modify content online from a central, shared interface.

Early CMS applications aimed to simplify the task of coding to streamline the website development process. As technological applications grew, the definition of a CMS received several interpretations. CMS systems today carry a multitude of functions and facilitate big data centralization, editing, publishing, and modification through a single back-end interface which also includes organization-level rules and processes that govern content creation and management. The latter are frequently viewed a part of the enterprise strategic decision-making and guide the selection, development, and implementation of a CMS. As content is produced, content managers can rely on the support of a CMS to create an infrastructure for multiple users to collaborate and contribute to all necessary knowledge management activities simultaneously.

A CMS may also provide tools for targeted online advertising, business-to-community communication, and audience engagement management.

Uses

Most frequently, examples of a CMS use include blogs, news sites, and shopping portals. In a nutshell, the CMS can keep the look and feel of a

website consistent while saving content managers the time and effort for creating new web pages as necessary, informing subscribers of the created content, or updating past content with new information. In a sense, a CMS can create standards for content management for small organizations and individual authors and make sure these standards are kept consistent.

A CMS can be used to manage the content of an externally focused organizational website or and internally targeted information sharing system. In either application, a CMS framework consists of two main elements: a content management application (CMA) and a content delivery application (CDA). The CMA functions as a tool for the content manager, who may not know HTML, to create, modify, and remove content from an online website independent of IT and webmaster support. Through the CDA, information is compiled and published online. Together, these elements create a central interface and make it possible for nontechnical users to add and edit text, control revisions, index data, and manage content for dissemination. In other words, a CMS allows a what-you-see-is-what-you-get editing and formatting by content authors who are not IT and web development specialists. A CMS, which may not require coding or direct management from an end user, provides more than content editing support. Robust systems may automatically generate navigation across the content, provide search functionality, facilitate indexing of content entries, track content activity by the system's users, and define user groups with varying security permissions and access to content.

In business applications, a CMS can be used for one-on-one marketing by delivering product information tailored to specific users' interest based on their past interaction with content. While corporations are the most frequent CMS users in their marketing efforts, a wide range of organizations, including nonprofits and public service organizations, can benefit from CMS use in relation to knowledge management and information dissemination efforts. Depending on the

size of an organization and the volume of published content, a CMS can be used to define roles and create workflow procedures for the collaborators who are involved in the content management. The workflow may include manual steps and standard operating procedures, or set up as an automated cascade of actions triggered by one of the content managers. As a central repository of data, a CMS may include documents, videos, pictures, customer and collaborator contact information, or scientific data. A CMS, aside from storing and publishing information, can provide necessary content linkages to show how new entries fit in and enhance previously existing content.

Functionality

Applying a systems approach to CMS evaluation and identifying the features of an ideal CMS, the following functions have been posited as most desirable (Garvin 2011; Han 2004): 1) a robust framework to facilitate knowledge use by end users, 2) a stable access to and ability to share information with other enterprise-level information systems, 3) a strategic plan for maintaining and disseminating relevant internally created and external knowledge, 4) a strategy for managing indexing and metadata associated with the content, and 5) a solution to reuse created content effectively.

CMSs offer complex and powerful functions. Most frequently CMS features can be broken into three major application areas: data maintenance, user access management, and system modification. Data maintenance allows creating a uniform structure across the variety of shared content through standardization. Standardization, aside from ensuring uniform presentation of content, creates structure for the shared data and allows for more robust analysis of the data themselves and their use. Other examples of data maintenance include template automation, which standardizes content appearance; data versioning, which allows for effective updates; and overall data

management, which includes content publishing, editing temporary removal from public access, and final archiving.

User access features include permission control and collaboration management. Permissions control allows CMS administrators to create user groups and assign access privileges and rights as well as to define how users, if any, can contribute their own content to an organization's website. Collaboration management features allow for multiple users to work on the content and for administrators to set up a workflow to cycle shared content for necessary review and editing and to delegate tasks automatically to the assigned users.

Finally, system modification features include scalability and upgrades. Scalability features may include the ability to create microsites within one CMS installation or addition of modules and plug-ins that extend the functionality of the original CMS. Plug-ins and module functions may include additional search options, content syndication, or moderation of user-shared content. Upgrade features refer to the regular updates to the CMS in accordance with the most current web standards.

Existing CMSs can be divided into two groups. The first group is *proprietary or enterprise CMSs*, created and managed by one particular developer with administrative access to edit and customize the website that uses the CMS. The second group is *open source CMSs*, which are open to any number of administrators who can make changes using any device. Although not necessary for immediate use, open-source CMSs allow programmers to adapt and modify the code of a system to tailor a CMS for the needs of an organization. The two CMS groups also differ in their approaches to the management of data and workflow. The first system frequently establishes standard operating procedures for content creation, review, and publishing, while the second usually lacks strict standardization. The fact that a CMS is open source does not mean that it cannot be used for enterprise-level content management. For example, Drupal and Wordpress can support a

broad range of content management needs, from a small blog to an enterprise-level application with complex access permissions and asynchronous, multiuser content publishing and editing.

CMS Options

A number of CMS systems are available, some of which are indicated in the brief descriptions and classifications below delineating some differences found in CMS products.

Blogs are CMS systems that are most appropriate for central content creation and controlled dissemination management. Most modern blog systems, like WordPress or Tumblr, require little initial HTML knowledge, but more advanced features may require application and use of scripts.

Online wikis are CMS that frequently crowd-source content and are usually edited by a number of users. Although the presentation of the content is usually static, such CMS systems benefit from the functionality that provides community members to add and edit content without coordination with a central knowledge repository.

Forums are dynamic CMS systems that provide community members with active conversation and discussion functionality, e.g., vBulletin or bbPress. Most forum systems are based on PHP and MySQL, but some online forum systems can be initiated without database or scripting knowledge.

Portals are another type of a CMS that can include both static and interactive content management features. Most portals are comprehensive and include wikis, forums, news feeds, etc. Some projects that support portal solutions are Joomla, Drupal, and Xoops, which support the development of portal sites in progressive, modular manner.

On the one hand, CMS offers tools for the collection, storage, analysis, and use of large amounts of data. On the other hand, big data are used to assess CMS measures and outcomes and to explore the relationships between them.

Cross-References

- [Business-to-Community \(B2C\)](#)
- [Content Moderation](#)
- [Semantic/Content Analysis/Natural Language Processing](#)
- [Sentiment Analysis](#)
- [Social Media](#)

Further Reading

- Barker, D. (2015). *Web content management: Systems, features, and best practices*. Sebastopol, CA: O'Reilly Media.
- Frick, T., & Eyler-Werve, K. (2015). *Return on engagement: content strategy and web design techniques for digital marketing*. Burlington, MA: Focal Press.
- Garvin, P. (2011). *Government information management in the 21st century: International perspectives*. Farnham/Surrey: Ashgate Pub.
- Han, Y. (2004). Digital content management: The search for a content management system. *Library Hi Tech*, v.22.

Content Moderation

Sarah T. Roberts

Department of Information Studies, University of California, Los Angeles, Los Angeles, CA, USA

Synonyms

[Community management](#); [Community moderation](#); [Content screening](#)

Definition

Content moderation is the organized practice of screening user-generated content (UGC) posted to Internet sites, social media, and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction. The process can result in UGC being

removed by a moderator, acting as an agent of the platform or site in question. Increasingly, social media platforms rely on massive quantities of UGC data to populate them and to drive user engagement; with that increase has come the concomitant need for platforms and sites to enforce their rules and relevant or applicable laws, as the posting of inappropriate content is considered a major source of liability.

The style of moderation can vary from site to site, and from platform to platform, as rules around what UGC is allowed are often set at a site or platform level and reflect that platform's brand and reputation, its tolerance for risk, and the type of user engagement it wishes to attract. In some cases, content moderation may take place in haphazard, disorganized, or inconsistent ways; in others, content moderation is a highly organized, routinized, and specific process. Content moderation may be undertaken by volunteers or, increasingly, in a commercial context by individuals or firms who receive remuneration for their services. The latter practice is known as commercial content moderation, or CCM. The firms who own social media sites and platforms that solicit UGC employ content moderation as a means to protect the firm from liability and negative publicity and to curate and control user experience.

History

The Internet and its many underlying technologies are highly codified and protocol-reliant spaces with regard to how data are transmitted within it (Galloway 2006), yet the subject matter and nature of content itself has historically enjoyed a much greater freedom. Indeed, a central claim to the early promise of the Internet as espoused by many of its proponents was that it was highly resistant, as a foundational part of both its architecture and ethos, to censorship of any kind.

Nevertheless, various forms of content moderation occurred in early online communities. Such content moderation was frequently undertaken by volunteers and was typically based on the

enforcement of local rules of engagement around community norms and user behavior. Moderation practices and style therefore developed locally among communities and their participants and could inform the flavor of a given community, from the highly rule-bound to the anarchic: the Bay Area-based online community the WELL famously banned only three users in its first 6 years of existence, and then only temporarily (Turner 2005, p. 499).

In social communities, on the early text-based Internet, mechanisms to enact moderation were often direct and visible to the user and could include demanding that a user alters a contribution to eliminate offensive or insulting material, the deletion or removal of posts, the banning of users (by username or IP address), the use of text filters to disallow posting of specific types of words or content, and other overt moderation actions. Examples of sites of this sort of content moderation include many Usenet groups, BBSes, MUDs, listservs, and various early commercial services.

Motives for people participating in voluntary moderation activities varied. In some cases, users carried out content moderation duties for prestige, status, or altruistic purposes (i.e., for the betterment of the community); in others, moderators received non-monetary compensation, such as free or reduced-fee access to online services, e.g., AOL (Postigo 2003). The voluntary model of content moderation persists today in many online communities and platforms; one such high-profile site where volunteer content moderation is used exclusively to control site content is Wikipedia.

As the Internet has grown into large-scale adoption and a massive economic engine, the desire for major mainstream platforms to control the UGC that they host and disseminate has also grown exponentially. Early on in the proliferation of so-called Web 2.0 sites, newspapers and other news media outlets, in particular, began noticing a significant problem with their online comments areas, which often devolved into unreadable spaces filled with invective, racist and sexist diatribes, name-calling, and irrelevant postings. These media firms began to employ a variety of

techniques to combat what they viewed as the misappropriation of the comments spaces, using in-house moderators, turning to firms that specialized in the large-scale management of such interactive areas and deploying technological interventions such as word filter lists or disallowing anonymous posting, to bring the comments sections under control. Some media outlets went the opposite way, preferring instead to close their comments sections altogether.

Commercial Content Moderation and the Contemporary Social Media Landscape

The battle with text-based comments was just the beginning of a much larger issue. The rise of Friendster, MySpace, and other social media applications in the early part of the twenty-first century has given way to more persistent social media platforms of enormous scale and reach. As of the second quarter of 2016, Facebook alone approached two billion users worldwide, all of whom generate content by virtue of their participation on the platform. YouTube reported receiving upwards of 100 hours of UGC video per minute as of 2014.

The contemporary social media landscape is therefore characterized by vast amounts of UGC uploads made by billions of users to massively popular commercial Internet sites and social media platforms with a global reach. Mainstream platforms, often owned by publicly traded firms responsible to shareholders, simply cannot afford the risk – legal, financial, and to reputation – that unchecked UGC could cause. Yet, contending with the staggering amounts of transmitted data from users to platforms is not a task that can currently be addressed reliably and at large scale by computers. Indeed, making nuanced decisions about what UGC is acceptable and what is not currently exceeds the abilities of machine-driven processes, save for the application of some algorithmically informed filters or bit-for-bit or hash value matching, which occur at relatively low levels of computational complexity.

The need for adjudication of UGC – video- and image-based content, in particular – therefore calls on human actors who rely upon their own linguistic and cultural knowledge and competencies to make decisions about UGC’s appropriateness for a given site or platform. Specifically, “they must be experts in matters of taste of the site’s presumed audience, have cultural knowledge about location of origin of the platform and of the audience (both of which may be very far removed, geographically and culturally, from where the screening is taking place), have linguistic competency in the language of the UGC (that may be a learned or second language for the content moderator), be steeped in the relevant laws governing the site’s location of origin and be experts in the user guidelines and other platform-level specifics concerning what is and is not allowed” (Roberts 2016). These human workers are the people who make up the legions of commercial content moderators: moderators who work in an organized way, for pay, on behalf of the world’s largest social media firms, apps, and websites who solicit UGC.

CCM processes may take place prior to material being submitted for inclusion or distribution on a site, or they may take place after material has already been uploaded, particularly on high-volume sites. Specifically, content moderation may be triggered as the result of complaints about material from site moderators or other site administrators, from external parties (e.g., companies alleging misappropriation of material they own; from law enforcement; from government actors) or from other users themselves who are disturbed or concerned by what they have seen and then invoke protocols or mechanisms on a site, such as the “flagging” of content, to prompt a review by moderators (Crawford and Gillespie 2016). In this regard, moderation practices are often uneven, and the removal of UGC may reasonably be likened to censorship, particularly when it is undertaken in order to suppress speech, political opinions, or other expressions that threaten the status quo.

CCM workers are called upon to match and adjudicate volumes of content, typically at rapid speed, against the specific rules or community

guidelines of the platform for which they labor. They must also be aware of the laws and statutes that may govern the geographic or national location from where the content emanates, for which the content is destined, and for where the platform or site is located – all of which may be distinct places in the world. They must be aware of the platform’s tolerance for risk, as well as the expectations of the platform for whether or how CCM workers should make their presence known.

In many cases, CCM workers may work at organizational arm’s length from the platforms they moderate. Some labor arrangements in CCM have workers located at great distances from the headquarters of the platforms for which they are responsible, in places such as the Philippines and India. The workers may be structurally removed from those firms, as well, via outsourcing companies who take on CCM contracts and then hire the workers under their auspices, in call center (often called BPO, or business process outsourcing) environments. Such outsourcing firms may also recruit CCM workers using digital piecework sites such as Amazon Mechanical Turk or Upwork, in which the relationships between the social media firms, the outsourcing company, and the CCM worker can be as ephemeral as one review.

Even when CCM workers are located on-site at a headquarters of a social media firm, they often are brought on as contract laborers and are not afforded the full status, or pay, of a regular full-time employee. In this regard, CCM work, wherever it takes place in the world and by whatever name, often shares the characteristic of being relatively low wage and low status as compared to other jobs in tech. These arrangements of institutional and geographic removal can pose a risk for workers, who can be exposed to disturbing and shocking material as a condition of their CCM work but can be a benefit to the social media firms who require their labor, as they can distance themselves from the impact of the CCM work on the workers. Further, the working conditions, practices, and existence of CCM workers in social media are little known to the general public, a fact that is often by design. CCM workers are frequently compelled to sign NDAs, or

nondisclosure agreements, that preclude them from discussing the work that they do or the conditions in which they do it. While social media firms often gesture at the need to maintain secrecy surrounding the exact nature of their moderation practices and the mechanisms they used to undertake them, claiming the possibility of users' being able to game the system and beat the rules if armed with such knowledge, the net result is that CCM workers labor in secret. The conditions of their work – its pace, the nature of the content they screen, the volume of material to be reviewed, and the secrecy – can lead to feelings of isolation, burnout, and depression among some CCM workers. Such feelings can be enhanced by the fact that few people know such work exists, assuming, if they think of it at all, that algorithmically driven computer programs take care of social media's moderation needs. It is a misconception that the industry has been slow to correct.

Conclusion

Despite claims and conventional wisdom to the contrary, content moderation has likely always existed in some form on the social Internet. As the Internet's many social media platforms grow and their financial, political, and social stakes increase, the undertaking of organized control of user expression through such practices as CCM will likewise only increase. Nevertheless, CCM remains a little discussed and little acknowledged aspect of the social media production chain, despite its mission-critical status in almost every case in which it is employed. The existence of a globalized CCM workforce abuts many difficult, existential questions about the nature of the Internet itself and the principles that have long been thought to undergird it, particularly, the free expression and circulation of material, thought, and ideas. These questions are further complicated by the pressures related to contested notions of jurisdiction, borders, application and enforcement of laws, social norms, and mores that frequently vary and often are in conflict with each other. The acknowledgement and understanding of the history of content moderation and the contemporary reality of large-scale

CCM is central to many of these core questions of what the Internet has been, is now, and will be in the future, and yet the continued invisibility and lack of acknowledgment of CCM workers by the firms for which their labor is essential means that such questions cannot fully be addressed. Nevertheless, discussions of moderation practices and the people who undertake them are essential to the end of more robust, nuanced understandings of the state of the contemporary Internet and to better policy and governance based on those understandings.

Cross-References

- [Algorithm](#)
- [Facebook](#)
- [Social Media](#)
- [Wikipedia](#)

Further Reading

- Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410–428.
- Galloway, A. R. (2006). *Protocol: How control exists after decentralization*. Cambridge, MA: MIT Press.
- Postigo, H. (2003). Emerging sources of labor on the internet: The case of America online volunteers. *International Review of Social History*, 48(S11), 205–223.
- Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. Tynes (Eds.), *The intersectional internet: Race, sex, class and culture online* (pp. 147–160). New York: Peter Lang.
- Turner, F. (2005). Where the counterculture met the new economy: The WELL and the origins of virtual community. *Technology and Culture*, 46(3), 485–512.

Content Screening

- [Content Moderation](#)

Context

- [Contexts](#)

Contexts

Feras A. Batarseh
College of Science, George Mason University,
Fairfax, VA, USA

Synonyms

Context; Contextual inquiry; Ethnographic observation

Definition

Contexts refer to all the information available to a software system that characterizes the situation it is running within. Context can be found across all types of software systems (where it is usually *intentionally* injected); however, it is mostly contained by intelligent systems. Intelligent systems are driven by two main parts, the intelligent algorithm and the data. More data means better understanding of context; therefore, Big Data can be a major catalyst in increasing the level of systems' self-awareness (i.e., the context they are operating within).

Contextual Reasoning

Humans have the ability to perform the processes of reasoning, thinking, and planning effectively; ideas could be managed, organized, and even conveyed in a comprehensible and quick manner. Context awareness is a “trivial” *skill* for humans. That is because humans receive years of training while observing the world around them, use agreed-upon syntax (language), comprehend the context in which they are in, and accommodate their understanding of events accordingly. Unfortunately, the same cannot be said about computers; this “understanding of context” is a major Artificial Intelligence (AI) challenge – if not the most important one in this age of technological transformations. With the latest massive diffusion of many new technologies such as smart mobile

phones, Big Data, tablets, and the cloud, AI applications such as context-aware software systems are gaining much traction. Context-aware systems have the advantage of dynamically adapting to current events and occurrences in the system and its surroundings. One of the main characteristics of such systems is to adjust the behavior of the system without human-user intervention (Batarseh 2014).

Recent applications of context include: (1) intelligent user interfaces' design and development, (2) context in software development, (3) robotics, and (4) intelligent software agents, among many others.

For context to reach its intended goals, it must be studied from both the technical and non-technical perspectives. Highlighting human aspects in AI (through context) will reduce the fears that many critics and researchers have toward AI. The arguments against AI have been mostly driven by the complexity of the human brain that is characterized by psychology, philosophy, and biology. Such arguments – many scientists believe – could be tackled by context, while context could be heavily improved by leveraging Big Data.

Conclusion

From the early days of AI research, many argued against the possibility of complete and general machine intelligence. One of the strongest arguments is that it is not yet clear how AI will be able to replicate the human brain and its biochemistry; therefore, it will be very difficult to represent feelings, thoughts, intuitions, moods, and awareness in a machine. Context, however, is a gateway to many of these very challenging aspects of intelligence.

Further Reading

Batarseh, F. (2014). Chapter 3: Context-driven testing. In *Context in computing: A cross-disciplinary approach for modeling the real world*. Springer. ISBN: 978-1-4939-1886-7.

Contextual Inquiry

► Contexts

Control Panel

► Dashboard

Core Curriculum Issues (Big Data Research/Analysis)

Rochelle E. Tractenberg
Collaborative for Research on Outcomes
and – Metrics, Washington, DC, USA
Departments of Neurology; Biostatistics,
Bioinformatics & Biomathematics; and
Rehabilitation Medicine, Georgetown University,
Washington, DC, USA

Definition

A curriculum is defined as the material and content that comprises a course of study within a school or college, i.e., a formal teaching program. The construct of “education” is differentiated from “training” based on the existence of a curriculum, through which a learner must progress in an evaluable, or at least verifiable, way. In this sense, a fundamental issue about a “big data curriculum” is what exactly is meant by the expression. “Big data” is actually not a sufficiently concrete construct to support a curriculum, nor even the integration of one or more courses into an existing curriculum. Therefore, the principal “core curriculum issue” for teaching and learning around big data is to articulate exactly what knowledge, skills, and abilities are to be taught and practiced through the curriculum. A second core issue is how to appropriately integrate those key knowledge, skills, and abilities (KSAs) into

the curricula of those who will *not* obtain degrees or certificates in disciplines related to big data – but for whom training or education in these KSAs is still desired or intended. A third core issue is how to construct the curriculum – whether the degree is directly related to big data or some key KSAs relating to big data are proposed for integration into another curriculum – in such a way that it is evaluable. Since the technical attributes of big data and its management and analysis are evolving nearly constantly, any curriculum developed to teach about big data must be evaluated periodically (e.g., annually) to ensure that what is being taught is relevant; this suggests that core underpinning constructs must be identified so that learners in every context can be encouraged to adapt to new knowledge rather than requiring retraining or reeducation.

Role of the Curriculum in “Education” Versus “Training”

Education can be differentiated from training by the existence of a curriculum in the former and its absence in the latter. The Oxford English Dictionary defines education as “the process of educating or being educated, the theory and practice of teaching,” whereas training is defined as “teaching a particular skill or type of behavior through regular practice and instruction.” The United Nations Educational, Scientific and Cultural Organization (UNESCO) highlights the fact that there may be an articulated curriculum (“intended”) but the curriculum that is actually delivered (“implemented”) may differ from what was intended. There are also the “actual” curriculum, representing what students learn, and the “hidden” curriculum, which comprises all the bias and *unintended* learning that any given curriculum achieves (<http://www.unesco.org/new/en/education/themes/strengthening-education-systems/quality-framework/technical-notes/different-meaning-of-curriculum/>). These types of curricula are also described by the Netherlands Institute for Curriculum Development (SLO, <http://international.slo.nl/>) and worldwide in

multiple books and publications on curriculum development and evaluation.

When a curriculum is being developed or evaluated with respect to its potential to teach about big data, each of these dimensions of that curriculum (intended, implemented, actual, hidden) must be considered. These features, well known to instructors and educators who receive formal training to engage in the kindergarten–12th grade (US) or preschool/primary/secondary (UK/Europe) education, are less well known among instructors in tertiary/higher education settings whose training is in other domains – even if their main job will be to teach undergraduate, graduate, postgraduate, and professional students. It may be helpful, in the consideration of curricular elements around big data, for those in the secondary education/college/university setting to consider what attributes characterize the curricula that their incoming students have experienced relating to the same content or topics.

Many modern researchers in the learning domains reserve the term “training” to mean “vocational training.” For example, Gibbs et al. (2004) identify training as specifically “skills acquisition” to be differentiated from instruction (“information acquisition”); together with socialization and the development of thinking and problem-solving skills, this information acquisition is the foundation of education overall. The vocational training is defined as a function of skills or behaviors to be learned (“acquired”) by practice in situ. When considering big data trainees, defined as individuals who participate in any training around big data that is outside of a formal curriculum, it is important to understand that there is no uniform cognitive schema, nor other contextual support, that the formal curriculum typically provides. Thus, it can be helpful to consider “training in big data” as appropriate for those who have completed a formal curriculum in data-related domains. Otherwise, skills that are acquired in such training, intended for deployment currently and specifically, may actually *limit* the trainees’ abilities to adapt to new knowledge, and thereby, lead to a requirement for retraining or reeducation.

Determining the Knowledge, Skills, and Abilities Relating to Big Data That Should Be Taught

The principal core curricular issue for teaching and learning around big data is to articulate exactly what knowledge, skills, and abilities are to be taught and practiced through the curriculum. As big data has become an increasingly popular construct (since about 2010), different stakeholders in the education enterprise have articulated curricular objectives in computer science, statistics, mathematics, and bioinformatics for undergraduate (e.g., De Veaux et al. 2017) and graduate students (e.g., Greene et al. 2016). These stakeholders include longstanding national or international professional associations and new groups seeking to establish either their own credibility or to define the niche in “big data” where they plan to operate. However, “big data” is not a specific domain that is recognized or recognizable; it has been described as a phenomenon (Boyd and Crawford 2012) and is widely considered *not* to be a domain for training or education on its own. Instead, knowledge, skills, and abilities relating to big data are conceptualized as belonging to the discipline of *data science*; this discipline is considered as existing at the intersection of mathematics, computer science, and statistics. This is practically implemented as the articulation of foundational aspects of each of these disciplines together with their formal and purposeful integration into a formal curriculum.

With respect to *data science*, then, generally, there is agreement that students must develop abilities to reason with data and to adapt to a changing environment, or changing characteristics of data (preferably both). However, there is not agreement on how to achieve these abilities. Moreover, because existing undergraduate course requirements are complex and tend to be comprehensive for “general education” as well as for the content making up a baccalaureate, associate, or other terminal degree in the postsecondary context, in some cases just a single course may be considered for incorporation into either required

or elective course lists. This would represent the *least* coherent integration of big data into a college/university undergraduate curriculum. In the construction of a program that would award a certificate, minor or major, if it seeks to successfully prepare students for work in or with big data, or statistics and data science, or analytics, or of other programs intended to train or prepare people for jobs that either focus on, or simply “know about,” big data must follow the same curricular design principles that every formal educational enterprise should follow. If they do not, they risk underperforming on their advertising and promises.

It is important to consider the role of *training* in the development, or consideration of development, of curricula that feature big data. In addition to the creation of undergraduate degrees and minors, Master’s degrees, post-baccalaureate certificate programs, and doctoral programs, all of which must be characterized by the curricula they are defined and created to deliver, many other “training” opportunities and workforce development initiatives also exist. These are being developed in corporate and other human resource-oriented domains, as well as in more open (open access) contexts. Unlike traditional degree programs, training and education around big data are unlikely to be situated specifically within a single disciplinary context – at least not exclusively. People who have specific skills, or who have created specific tools, often create free or easily accessible representations of the skills or tool – e.g., instructional videos on YouTube or as formal courses of varying lengths that can be read (slides, documentation) or watched as webinars. Examples can be found online at sites including Big Data University (bigdatauniversity.com), created by IBM and freely available, and Coursera (coursera.org) which offers data science, analytics, and statistics courses as well as eight different specializations, comprising curated series of courses – but also many other topics. Coursera has evolved many different educational opportunities and some curated sequences that can be completed to achieve “certification,” with different costs depending on the extent of student engagement/commitment. The Open University (www.open.ac.uk) is essentially an

online version of regular university courses and curricula (and so is closer to “education” than “training”) – degree and certificate programs all have costs associated and also can be considered to follow a formal curriculum to a greater extent than any other option for widely accessible training/learning around big data. These examples represent a continuum that can be characterized by the attention to the curricular structure from minimal (Big Data University) to complete (The Open University). The individual who selects a given training opportunity, as well as those who propose and develop training programs, must articulate exactly what knowledge, skills, and abilities are to be taught and practiced. The challenge for individuals making selections is to determine how *correctly* an instructor or program developer has described the achievements the training is intended to provide. The challenge for those curating or creating programs of study is to ensure that the learning objectives of the curriculum are met, i.e., that the *actual* curriculum is as high a match to the *intended* curriculum as possible. Basic principles of curriculum design can be brought to bear for acceptable results in this matching challenge. The stronger the adherence to these basic principles, the more likely a robust and evaluable curriculum, with demonstrable impact, will result. This is not specific to education around big data, but with all the current interest in data and data science, these challenges rise to the level of “core curriculum issues” for this domain.

Utility of Training Versus a Curriculum Around Big Data

De Veaux et al. (2017) convened a consensus panel to determine the fundamental requirements for an undergraduate curriculum in “data science.” They articulated that the main topical areas that comprise – and must be leveraged for appropriate baccalaureate-level training in – this domain are as follows: data description and curation, mathematical foundations, computational thinking, statistical thinking, data modeling, communication, reproducibility, and ethics. Since computational and statistical thinking, as well as data modeling, all require somewhat

different mathematical foundations, this list shows clearly the challenges in selecting specific “training opportunities” to support development of new skills in “big data” for those who are not already trained in quantitative sciences to at least some extent. Moreover, arguments are arising in many quarters (science and society, philosophy/ethics/bioethics, and professional associations like the Royal Statistical Society, American Statistical Association, and Association of Computing Machinery) that “ethics” is not a single entity but, with respect to big data and data science, is a complex – and necessary – type of reasoning that cannot be developed in a single course or training opportunity. The complexity of reasoning that is required for competent work in the domain referred to exchangeably as “data analytics,” “data science,” and “big data”, which includes this ability to reason ethically, underscores the point that piecemeal training will be unsuccessful unless the trainee possesses the ability to organize the new material together with extant (high level) reasoning abilities, or at least a cognitive/mental schema within which the diverse training experiences can be integrated for a comprehensive understanding of the domain.

However, the proliferation of training opportunities around big data suggests a pervasive sense that a formal curriculum is not actually needed – just *training* is. This may arise from a sense that the technology is changing too fast to create a whole curriculum around it. Training opportunity creators are typically experts in the domain, but may not necessarily be sufficiently expert in teaching and learning theories, or the domains from which trainees are coming, to successfully translate their expertise into effective “training.” This may lead to the development of new training opportunities that appear to be relevant, but which can actually contribute only minimally to an individual trainee’s ability to function competently in a new domain like big data, because they do not also include or provide contextualization or schematic links with prior knowledge.

An example of this problem is the creation of “competencies” by subject matter expert consensus committees, which are then used to create “learning plans” or checklists. The subject matter

experts undoubtedly can articulate what competencies are required for functional status in their domain. However, (a) a training experience developed to fill in a slot within a competency checklist often fails to support teaching and learning around the *integration* of the competencies into regular practice; and (b) curricula created in alignment with competencies often do not promote the actual development *and refinement* of these competencies. Instead, they may tend to favor the checking-off of “achievement of competency X” from the list.

Another potential challenge arises from the opposite side of the problem, learner-driven training development. “What learners want and need from training” *should* be considered together with what experts who are actually using the target knowledge, skills, and abilities believe learners need from training. However, the typical trainee will not be sufficiently knowledgeable to choose the training that is *in fact* most appropriate for their current skills and learning objectives. The construct of “deliberate practice” is instructive here. In their 2007 Harvard Business Review article, “The making of an expert,” Ericsson, Prietula, and Cokely summarize Ericsson’s prior work on expertise and its acquisition, commenting that “(y)ou need a particular kind of practice – *deliberate practice* - to develop expertise” (emphasis in original, p. 3). *Deliberate practice* is practice where weaknesses are specifically *identified and targeted* – usually by an expert both in the target skillset and perhaps more particularly in identifying and remediating specific weaknesses. If a trainee is not (yet) an expert, determining how best to address a weakness that one has self-identified can be another limitation on the success of a training opportunity, if it focuses on what the learner wants or believes they need without appeal to subject matter experts. This perspective argues for the incorporation of expert opinion into the development, descriptions, and contextualizations of training, i.e., the importance of deliberate practice in the assurance that as much as possible of the intended curriculum becomes the *actual* curriculum. Training opportunities around big data can be developed to support, or fill in gaps, in a formal curriculum; without this context, training in big data may not be as successful as desired.

Conclusions

A curriculum is a formal program of study, and basic curriculum development principles are essential for effective education in big data – as in any other domains. Knowledge, skills, and abilities, and the levels to which these will be both developed and integrated, must be articulated in order to structure a curriculum to optimize the match between the intended and the actual curricula. The principal core curricular issue for teaching and learning around big data is to articulate exactly what knowledge, skills, and abilities are to be taught and practiced. A second core issue is that the “big data” knowledge, skills, and abilities may require more foundational support for training of those who will not obtain, or have not obtained, degrees or certificates in disciplines related to big data. A third core issue is how to construct the curriculum in such a way that the alignment of the intended and the actual objectives is evaluable and modifiable as appropriate. Since the technical attributes of big data and its management and analysis are evolving nearly constantly, any curriculum developed to teach about big data must be evaluated periodically to ensure the relevance of the content; however the alignment of the intended and actual curricula must also be regularly evaluated to ensure learning objectives are achieved and achievable.

Further Reading

- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society*, 15(5), 662–679.
- De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., et al. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and its Applications*, 4, 2.1–2.16. <https://doi.org/10.1146/annurev-statistics-060116-053930>. Downloaded from <http://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf>. 2 Jan 2017.
- Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review* 85(7–8):114–121, 193. Downloaded from <https://hbr.org/2007/07/the-making-of-an-expert>. 5 June 2010.
- Gibbs, T., Brigden, D., & Hellenberg, D. (2004). The education versus training and the skills versus competency debate. *South African Family Practice*, 46(10), 5–6. <https://doi.org/10.1080/20786204.2004.10873146>.
- Greene, A. C., Giffin, K. A., Greene, C. S., & Moore, J. H. (2016). Adapting bioinformatics curricula for big data. *Briefings in Bioinformatics*, 17(1), 43–50. <https://doi.org/10.1093/bib/bbv018>.

Corporate Social Responsibility

Yon Jung Choi

Center for Science, Technology, and Innovation Policy, George Mason University, Fairfax, VA, USA

Big data has become a popular source for businesses to analyze various aspects of human psychology and behaviors and organizational processes and practices. However, the growing corporate use of big data has raised several ethical and social concerns, especially related to the use of personal data for commercial interests and possible infringement of fundamental rights (e.g., privacy) (Herschel and Miori 2017; Flyverbom et al. 2019). These concerns are primarily issues of corporate social responsibility (CSR), referring in general to the responsibilities of businesses for their impacts on society, including economic, legal, environmental, and social responsibilities. Despite growing public concerns, practical efforts to connect big data with CSR have been rarely made (Richards and King 2014; Zwitter 2014; Napier 2019). Corporate use of big data poses both risks and opportunities for society. The following sections summarize various CSR-related concerns and possible contributions to CSR brought by corporate use of big data that have been identified by scholars and practitioners.

CSR-Related Concerns

There are several social and ethical concerns in the use of big data by businesses. The following are

some of the issues that have been identified by scholars, the media, and the public.

- **Infringement of privacy:** The most well-publicized concern is the issue of a possible violation of privacy. A vast amount of personal data, including personal information, interests, movement, and relationships, are in the hands of internet companies that can be easily approached and exchanged among them with or without users' consent (Zwitter 2014; Herschel and Miori 2017; Flyverbom et al. 2019). Thus, user's privacy can be easily infringed upon by these companies under the current system, which lacks sufficient regulations.
- **Transparency and consumer rights:** In general, corporate information about how to create, manage, exchange, and protect big data composed of users' information is not open to the public because it is considered a proprietary asset. Consumers have little knowledge of how their information is gathered and handled by companies. Users' data that companies can access is extensive, including geolocation, photos and videos, contacts, text messages, and emails, and users are not fully aware of this (Flyverbom et al. 2019, p. 7). This raises not only concerns over privacy but also over consumers' "right-to-know" about the risks to their lives and well-being that can be caused by this.
- **Data monopoly and manipulation of individual desires and needs:** Because many internet companies enjoy exclusive rights to the big data generated by their platforms, there is an issue of data monopoly, as evidenced by the antitrust lawsuits filed against Facebook (Iyengar 2020; Flyverbom et al. 2019; Zwitter 2014). Scholars warn of internet companies' ability to shape "our views of the world by managing, editing, and controlling information in ways that have important consequences for individuals, organizations, and societies alike" (Flyverbom et al. 2019, p. 8; see also Flyverbom 2016; Helveston 2016). For example, how individuals represent themselves in the digital world has increasingly

depended on decisions made by a handful of social network service (SNS) companies. People's public images and behaviors are greatly influenced by how these companies guide people through their digital platforms.

- **Politicization and commodification of personal information:** Another growing concern over big data handled by internet companies is the possibility of political use for mass surveillance and commodification of personal information, as evidenced by Edward Snowden's revelations about the US National Security Agency's mass surveillance, and the growing suspicion of surveillance by the Chinese government of online information provided by internet companies (Bauman et al. 2014; Hou 2017). Companies have the option to sell or provide their big data to governments and/or private companies for political or commercial purposes, which raises serious ethical concerns (Flyverbom et al. 2019).

Scholars have pointed out that laws and regulations over these issues are lacking at both national and global levels, and the lives and well-being of the public are significantly at stake. Some argue that companies, as "socially responsible" actors, should consider the ethical management of big data as part of their business account (Fernandes 2018; Flyverbom et al. 2019). In other words, socially conscious management of big data is argued to be part of the main areas of CSR and should be scrutinized by the public through sources such as CSR/transparency reports, especially for those companies creating and dealing with big data. In this regard, the main areas of CSR that are widely recognized, especially in academia and industry, are environmental protection, employee welfare, stakeholder involvement, anticorruption, human rights protection, and community development (UNGC 2020; GRI 2018). Others also insist on the necessity of developing more inclusive decision-making mechanisms either within corporate governance or collaboratively by inviting various stakeholders, enabling them to serve their interests more adequately (Flyverbom 2016; Flyverbom et al. 2019).

Implications of Big Data Contributions to CSR

Scholars and practitioners are increasingly engaging big data in relation to CSR, as summarized below:

- **Measurement, Assessment, and Enhancement of CSR performance:** Big data can be used to measure and evaluate CSR and sustainable development activities of companies by analyzing environmental/social data and communications of corporate stakeholders (Barbeito-Caamaño and Chalmeta 2020; Jeble et al. 2018). Big data analytics may help to ease the difficulty of measuring and evaluating intangible social values influenced by corporate practices (Jeble et al. 2018). In addition, new information technologies using big data can also help manage and enhance companies' social and environmental performances (Carberry et al. 2017; Akhtar et al. 2018; Napier 2019).
- **Better management of stakeholders:** Scholars and practitioners have recognized the potential of big data in stakeholder management. For instance, SAP Ariba, a software company, has developed procurement intelligence (a method of big data analytics) to identify “unethical or unsustainable business practices” of suppliers of companies and therefore enhance their supply chain management (York 2018). Big data analytics can also be used for better management of other stakeholders, such as employees and consumers, with a more in-depth understanding of their needs and preferences.
- **Contributions to creating social goods:** Companies arguably can generate significant social benefits with a deeper understanding of the people, organizations, culture, and values of a society if they use and manage big data more responsibly and implement more socially conscious practices. More specifically, companies can make a significant contribution to generating public goods such as “improvements in health care, education, and urban planning” through big data

analytics (Napier 2019; Flyverbom et al. 2019, p. 12).

Although it is still at an early stage, the debate on corporate social responsibility surrounding big data has pointed out both risks and benefits to society. A better understanding is needed of the economic, political, environmental, and social impact and implications of corporate use of big data in order to discuss and establish the proper and ethical roles and responsibilities of business in society.

Further Reading

- Akhtar, P., Khan, Z., Frynas, J., Tse, Y., & Rao-Nicholson, R. (2018). Essential micro-foundations for contemporary business operations: Top management tangible competencies, relationship-based business networks and environmental sustainability. *British Journal of Management*, 29, 43–62.
- Barbeito-Caamaño, A., & Chalmeta, R. (2020). Using big data to evaluate corporate social responsibility and sustainable development practices. *Corporate Social Responsibility & Environmental Management*, 27(6), 2831–2848.
- Bauman, Z., Bigo, D., Esteves, P., Guild, E., Jabri, V., Lyon, D., & Walker, R. B. J. (2014). After Snowden: Rethinking the impact of surveillance. *International Political Sociology*, 8(2), 121–144.
- Carberry, E., Bharati, P., Levy, D., & Chaudhury, A. (2017). Social movements as catalysts for corporate social innovation: Environmental activism and the adoption of green information systems. *Business & Society*, 58(5), 1083–1127.
- Fernandes, K. (2018, November 2). CSR in the era of big data. *The CSR Journal*. <https://thecsrjournal.in/csr-era-big-data-analytics-private-companies/>.
- Flyverbom, M. (2016). Disclosing and concealing: Internet governance, information control, and the management of visibility. *Internet Policy Review*, 5(3), 1–15.
- Flyverbom, M., Deibert, R., & Matten, D. (2019). The governance of digital technology, big data, and the internet: New roles and responsibilities for business. *Business & Society*, 58(1), 3–19.
- Global Reporting Initiative (GRI). (2018). *GRI standards*. <https://www.globalreporting.org/media/55yhvety/gri-101-foundation-2016.pdf>.
- Helveston, M. (2016). Consumer protection in the age of big data. *Washington University Law Review*, 93(4–5), 859.
- Herschel, R., & Miori, V. (2017). Ethics & big data. *Technology in Society*, 49, 31–36.
- Hou, R. (2017). Neoliberal governance or digitalized autocracy? The rising market for online opinion

- surveillance in China. *Surveillance & Society*, 15(3/4), 418–424.
- Iyengar, R. (2020, December 11). The antitrust case against Facebook: Here's what you need to know. *CNN Business*. <https://www.cnn.com/2020/12/11/tech/facebook-antitrust-lawsuit-what-to-know/index.html>.
- Jebble, S., Dubey, R., Childe, S., Papadopoulos, T., Roubaud, D., & Prakash, A. (2018). Impact of big data and predictive analytics capability on supply chain sustainability. *International Journal of Logistics Management*, 29(2), 513–538.
- Napier, E. (2019). *Technology enabled social responsibility projects and an empirical test of CSR's impact on firm performance*. (Doctoral dissertation, Georgia State University). ScholarWorks @ Georgia State University https://scholarworks.gsu.edu/marketing_diss/50.
- Richards, N. M., & King, J. H. (2014). Big data ethics. *Wake Forest Law Review*, 49, 393–432.
- United Nations Global Compact (UNGC). (2020). *UNGC principles*. <https://www.unglobalcompact.org/what-is-gc/mission/principles>.
- York, M. (2018, March 26). Intelligence-driven CSR: Putting big data to good use. *COP Rising*. <https://cporising.com/2018/03/26/intelligence-driven-csr-putting-big-data-to-good-use/>.
- Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(3), 1–6.

Corpus Linguistics

Patrick Juola
Department of Mathematics and Computer
Science, McAnulty College and Graduate School
of Liberal Arts, Duquesne University, Pittsburgh,
PA, USA

Introduction

Corpus linguistics is, broadly speaking, the application of “big data” to the science of linguistics. Unlike traditional linguistic analysis [caricatured by Fillmore (1992) as “armchair linguistics”], which relies on native intuition and introspection, corpus linguists rely on large samples to quantitatively analyze the distribution of linguistic items. It has therefore tended to focus on what can be easily measured by computer and quantified, such as words, phrases, and word-based grammar, instead of more abstract concepts such as discourse or formal syntax. With the advent of

high-powered computers and the increased availability of machine-readable texts, it has become a major force in modern linguistic research.

History

The use of corpora for language analysis long predates computers. Theologians were making Biblical concordances in the eighteenth century, and Samuel Johnson started a tradition followed to this day (e.g., most famously by the *Oxford English Dictionary*) of compiling collections of quotations from prestigious literature to form the basis of his dictionary entries. Dialect dictionaries such as the *Dictionary of American Regional English (DARE)* are typically compiled on the basis of questionnaires or interviews of hundreds or thousands of people.

The first and possibly most famous use of computer-readable corpora was the million-word Brown corpus (Kučera and Nelson Francis 1967). The Brown corpus consists of 500 samples, each of about 2000 words, collected from writings published in the United States in 1961. Genre coverage includes nine categories of “informative prose” and six of “imaginative prose,” including inter alia selections of press reportage, learned journals, religious tracts, and mystery novels. For many years, the Brown corpus was the only large corpus available and the de facto standard. Even today, the Brown corpus has influenced the design and collection of many later corpora, including the LOB corpus (British English), the Kolhapur corpus (Indian English), the Australian Corpus of English, and the 100-million-word British National Corpus.

Improvements in computer science made three major innovations possible. First, more powerful computers made the actual task of processing data much easier and faster. Second, improvements in networking technology make it practical to distribute data more easily and even to provide corpus analysis as a service via web interfaces such as <https://books.google.com/ngrams> or <https://corpus.byu.edu/coca>. Finally, the development of online publishing via platforms such as the Web makes it much easier to collect data simply by

scraping databases; similarly, improvements in optical character recognition (OCR) technology have made large-scale scanning projects such as <https://www.google.com/googlebooks/about/> more practical.

Theory

From the outset, corpus linguistics received pushback from some theoretical linguists. Chomsky, for example, stated that “Corpus linguistics doesn’t mean anything.” [cited in McEnery and Hardy 2012]. Meyer (2002) describes a leading generative grammarian as saying “‘the only legitimate source of grammatical knowledge’ about a language [is] the intuitions of the native speaker.” Statistics often provide an inadequate explanatory basis for linguistic findings. For example, one can observe that the sentence **Studied for the exam* does not appear in a sample of English writing, but *He studied for the exam* does. It requires substantial intuition, ideally by a native speaker, to observe that the first form is not merely rare but actively ungrammatical. More specifically, intuition is what tells us that English (unlike Italian) generally requires that all sentences have an explicit subject. (Worse, the sentence *Studied for the exam* might appear, perhaps as an example, or perhaps in an elliptical context. This might suggest to the naïve scholar that the only difference between the two forms is how common each is.)

Similarly, just because a phenomenon is common does not make it important or interesting. Fillmore (1992) caricatures this argument as “if natural scientists felt it necessary to portion out their time and attention to phenomena on the basis of their abundance and distribution in the universe, almost all of the scientific community would have to devote itself exclusively to the study of interstellar dust.”

At the same time, intuitions are not necessarily reliable; perhaps more importantly, they are unshared. Fillmore (1992) cites as an example his theory that “the colloquial gesture-requiring *yea*, as in *It was about yea big*,” couldn’t be used in a context when the listener couldn’t see the speaker. However, Fillmore acknowledged that

people have been observed using this very expression over the telephone, indicating that his intuitions about acceptability (and the reasons for unacceptability) are not necessarily universally shared. Alternatively, people’s intuitions may not accurately reflect their actual use of language, a phenomenon found in other studies of human expertise. Observation of actual use can often be made only by using empirical, that is, corpus-type, evidence.

Corpora therefore provide observational evidence about the use of language – which patterns are used, and by extension, which might not be – without necessarily diving deeper into a description of the types of patterns used or an explanation of the underlying processes. Furthermore, they provide a way of capturing the effects of the intuitions of hundreds, thousands, or millions of people instead of a single researcher or small team. They enable investigation of rare phenomena that researchers may not have imagined and allow quantitative investigations with greater statistical power to discern subtle effects.

Applications of Corpus Linguistics

Corpora are used for many purposes, including language description and as a resource for language learning. One long-standing application is compiling dictionaries. By collecting a large enough number of samples of a specific word in use, scholars can identify the major categories of meanings or contexts. For example, the English word *risk* typically takes three different types of direct objects – you can “risk” an action (*I wouldn’t risk that climb*), what you might do as a consequence (*... because you risk a fall on the slippery rocks*), or even the consequence of the consequence, what you might lose (*... and you would risk your life*). The different meanings of polysemous terms (words with multiple meanings) like *bank* (the edge of a river, a financial institution, and possibly other meanings, such as *a bank shot*) can be identified from similar lists. Corpora can help identify frequently occurring patterns (collocations) such as idioms and can identify grammatical patterns such as the types

of grammatical structure associated with specific words. (For instance, the word *give* can take an indirect object, as in *John gave Mary a taco*, but *donate* typically does not – constructions such as **John donated the museum a statue* are vanishingly rare.)

Another application is in detecting illustrating language variation and change. For example, corpora of Early Modern English such as the ARCHER corpus can help illustrate differences between Early Modern and contemporary English. Similar analysis across genres can show different aspects of genre variations, such as what Biber (cited in McEnery and Hardy 2012) termed “narrative vs. non-narrative concerns,” a concept describing the relation of specific past events with specific people. Other studies show differences between groups or even between specific individuals (Juola 2006), a capacity of interest to law enforcement (who may want to know which specific individual was associated with a specific writing, like a ransom note).

Corpora can also provide source material to train statistical language processors on. Many natural language tasks (such as identifying all the proper nouns in a document or translating a document from one language to another) have proven to be difficult to formalize in a rule-based system. A suitable corpus (perhaps annotated by partial markup done by humans) can provide the basis instead for a machine learning system to determine complex statistical patterns associated with that task. Rather than requiring linguists to list the specific attributes of a proper noun or the specific rules governing the exact translation of the verb *to wear* into Japanese, the system can “learn” patterns associated with these distinctions and generalize them to novel contexts. Other examples of such natural language processing problems include parsing sentences to determine their constituent structure, resolving ambiguities such as polysemous terms, providing automatic markup such as tagging the words in a document for their parts of speech, answering client questions (“Siri, where is Intel based?”), or determining whether the overall sentiment expressed in an online review is favorable or unfavorable.

Conclusions

The use of “big data” in language greatly expands the types of research questions that can be addressed and provides valuable resources for use in large-scale language processing. Despite early criticisms, it has helped to establish many new research methods and findings and continues to be an important and now-mainstream part of linguistics research.

Cross-References

► [Google Books Ngrams](#)

Further Reading

- Fillmore, C. J. (1992). “Corpus linguistics” or “computer-aided armchair linguistics”. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel symposium 82. 4–8 August 1991* (pp. 35–60). Berlin: Mouton de Gruyter.
- Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.
- Kučera, H., & Nelson Francis, W. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.
- McEnery, T., & Hardy, A. (2012). *Corpus linguistics: Method, theory, practice*. Cambridge: Cambridge University Press.
- Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Correlation Versus Causation

R. Bruce Anderson^{1,2} and Matthew Geras²

¹Earth & Environment, Boston University,
Boston, MA, USA

²Florida Southern College, Lakeland, FL, USA

Both the terms correlation and causation are often used when interpreting, evaluating, and describing statistical data. While correlations and causations can be associated, they do not need to be

related or linked. A correlation and a causation are two distinct and separate statistical terms that can each individually be used to describe and interpret different types of data. Sometimes the two terms are mistakenly used interchangeably, which could misrepresent important trends in a given data set. The danger of using these terms as synonyms has become even more problematic in recent years with the continued emergence of research projects relying on big data. Any time a researcher utilizes a large dataset with thousands of observations, they are bound to find correlations between variables; however, with such large datasets, there is an inherent risk that these correlations are spurious opposed to causal.

A correlation, sometimes called an association, describes the linear relationship or lack of a linear relationship between two or more given variables. The purpose of measuring data sets for correlation is to determine the strength of association between different variables. There are several statistical methods to determine whether a correlation exists between variables, whether that correlation is positive or negative, and whether the correlation shows a strong association or a weak association. Correlations can be either positive or negative. A positive correlation occurs when one variable increases as another variable increases or when one variable decreases as another variable decreases. For a positive correlation to be evident, the variables being compared have to move in tandem with one another. A negative correlation behaves in the opposite pattern; as one variable increases another variable decreases or as the first variable decreases, the second variable increases. With negative correlations, the variables in question need to move in the opposite direction of one another. It is also possible for no correlation to exist between two or more different variables.

Statistically, correlations are stated by the correlation coefficient r . A correlation coefficient with a value that is greater than zero up to and including one indicates a positive linear correlation, where a score of one is a perfect positive correlation and a positive score close to zero represents variables that have a very weak or limited positive correlation. One example of a strong positive correlation would be the amount of time spent exercising and the amount of calories

burned off through the course of a workout. This is an example of positive correlation because as the amount of time spent exercising increases, so does the amount of calories being burned. A correlation coefficient value that ranges from anything less than zero to negative one indicates a negative linear correlation, where a score of negative one is a perfect negative correlation and a negative score close to zero represents variables that have a very weak or limited negative correlation. An example of a negative correlation would be the speed at which a car is traveling and the amount of time it takes that car to arrive at its destination. As the speed of the car decreases, the amount of time traveling to the destination increases. A correlation coefficient of zero indicates that there is no correlation between the variables in question. Additionally, when two variables result in a very small negative or positive correlation, such as -0.01 or 0.01 , the corresponding negative or positive correlation is often considered to have very little substantive meaning and thus variables in cases such as these are also often considered to have little to no correlation.

R or the correlation coefficient can be determined in several ways. One way to determine the correlation between two different variables is through the use of graphing methods. Scatterplots can be used to compare more than one variable. When using a scatterplot, one variable is graphed along the x-axis, while the other variable is graphed along the y-axis. Once all of the points are graphed, a line of best fit, a line running through the data points where half of the data points are positioned above the line and half of the data points are graphed below the line, can be used to determine whether there is a correlation between the variables being examined. If the line of best fit has a high positive slope, then there is a strong correlation between the variables and if the slope of the line of best fit is a high negative number, the correlation between the variables is negative. Finally, if the slope of the line of best fit is close to zero, or close to being a straight line, there is little to no correlation between the variables. A correlation coefficient can also be determined numerically by the use of Karl Pearson's coefficient of correlation formula. Using this formula involves taking Sigma,

or the sum, of all of the differences between each individual x values and the mean x value multiplied by the sum of all of the differences between each individual y values and the mean y value. This product then becomes the numerator of the equation and is divided by the product of N , or the number of observations, the standard deviation of x , and the standard deviation of y . Fortunately, due to the ever advancing and expanding field of technology, correlation coefficients can now be determined practically instantly through the use of technology such as graphing calculators and different types of statistical analysis software. Due to the speed and increasing availability of these types of software, the practice of manually calculating correlation coefficients is limited mostly to classrooms.

In relation to experimentation and data analysis, causation means that changes in one variable directly relate and influence changes in another variable. When causation or causality is present between two variables, the first variable, the variable that is the cause, may bring the second variable into occurrence or influence the direction and movement of the second variable. While on the surface or by reading definitions alone, correlation and causation may appear to be the same thing or at least that correlations prove causation, this is not the case. Correlations do not necessarily indicate causation because causation is not always straight forward and is difficult to prove. Even a strong correlation cannot immediately be considered causation due the possibility of confounding variables. Confounding variables are extraneous variables or variables that are not being controlled for or measured in a particular experiment or survey but could still have impact on the results. When examining variables x and y , it may be possible to determine that x and y have a positive correlation, but it is not as clear that x causes y because confounding variables w , z , etc. could also be influencing the outcome of y unbeknownst to the individuals examining the data. For example, the number of people at the beach could increase as the daily temperature increases, but it is not possible to know that the increase in temperature caused the increased beach attendance. Other variables could be in play; more people could be at the beach because it is a federal

holiday or because the local public swimming pool is closed for repairs. Due to the possibility, of confounding variables, it is not possible to determine causation from correlation alone.

Despite this, it is possible to estimate whether there is a causal relationship between two variables. One way to evaluate causation is through the use of experimentation with random samples. Through the use of either laboratory or field experiments, researchers may be able to estimate causation since they will be able to control or limit the effect of possible confounding variables. For the beach example listed above, researchers could measure beach attendance on a daily basis for a given period of time. This would help them to eliminate potential extraneous variables, such as holidays, because data will be collected according to a set plan and not just based on one day's worth of observations. Additionally, the coefficient of determination or r^2 can be used to measure whether two variables cause the changes in one another. The coefficient of determination is equivalent to the correlation coefficient multiplied by itself (squared). Since any number squared is a positive number, the coefficient of determination will always have a positive value. As is the case with positive correlations, the closer to one that a coefficient of determination is, the more likely it is that the first variable being examined caused the second variable. This is because r^2 tells what percent of the independent variable x explains what happened to the dependent variable. The closer r squared is to 1, the better x explains y . Of these two methods, experimental research is more widely accepted when claiming causality between two variables, but both methods provide better indicators of causality than does a single correlation coefficient. This is especially true when utilizing big data since the risk of finding spurious correlations increases as the number of observations increases.

Cross-References

- [Association Versus Causation](#)
- [Data Integrity](#)
- [Data Processing](#)
- [Transparency](#)

Further Reading

- Correlation Coefficients. (2005, January 1). Retrieved 14 Aug 2014, from <http://www.andrews.edu/~calkins/math/edrm611/edrm05.htm>.
- Green, N. (2012, January 6). Correlation is not causation. Retrieved 14 Aug 2014, from <http://www.theguardian.com/science/blog/2012/jan/06/correlation-causation>.
- Jaffe, A. (2010, January 1). Correlation, causation, and association – What does it all mean? Retrieved 14 Aug 2014, from <http://www.psychologytoday.com/blog/all-about-addiction/201003/correlation-causation-and-association-what-does-it-all-mean>.

COVID-19 Pandemic

Laurie A. Schintler
George Mason University, Fairfax, VA, USA

Overview

In 2020, COVID-19 took the world by storm. First discovered in China, the novel coronavirus quickly and aggressively spread throughout Asia and then to the rest of the world. As of November 2020, COVID-19 infections, deaths, and hospitalizations continue to rise with no end in sight. In attempts to manage the pandemic's progression, big data are playing an innovative and instrumental role (Lin and Hou 2020; Pham et al. 2020; Vaishya et al. 2020). Specifically, big data are being used for:

1. Disease surveillance and epidemiological modeling
2. Understanding disease risk factors and triggers
3. Diagnosis, treatment, and vaccine development
4. Resource optimization, allocation, and distribution
5. Formulation and evaluation of containment policies

Various sources of structured and unstructured big data, such as mobile phones, social media platforms, search engines, biometrics sensors, genomics repositories, images and videos,

electronic health records, wearable devices, satellites, wastewater systems, and scholarly articles and clinical studies, are being exploited for these purposes (Lin and Hou 2020).

Working hand-in-hand with big data is an integrated set of emerging digital, cyber-physical, and biological tools and technologies (Ting et al. 2020). Indeed, the COVID-19 pandemic has unfolded during a period of rapid, disruptive, and unprecedented technological change referred to as a Fourth Industrial Revolution. In this context, emerging technologies, such as Artificial Intelligence (AI) enabled by deep learning, have been invaluable in the fight against COVID-19, particularly for transforming big data into actionable insight (i.e., translating evidence to action). Other technologies such as blockchain, the Internet of Things (IoT), and smart mobile devices provide big data sources and the means for processing, storing, and vetting massive bits and bytes of information.

Benefits and Opportunities

Big data are helping to address the informational challenges of the COVID-19 pandemic in various ways. First, in a global disease outbreak like COVID-19, there is a need for timely information, especially given that conditions surrounding the disease's spread and understanding of the disease itself are very fluid. Big data tends to have a high velocity, streaming in at a relatively fast pace – in some cases, second-by-second. In fact, data are produced now at an exponentially higher speed than in other recent pandemics, e.g., the SARS 2002–2003 outbreak. In COVID-19, such fast-moving big data enable continuous surveillance of epidemiological dynamics and outcomes and forecasting and on-the-fly predictions and assessments, i.e., “nowcasting.” For example, big data produced by Internet and mobile phone users are helping with the ongoing evaluation of non-pharmaceutical interventions, such as shutdowns, travel bans, quarantines, and social distancing mandates (Oliver et al. 2020).

In a pandemic, there is also a dire need to understand the pathology of and risk factors

behind the disease in question, particularly for the rapid development and discovery of effective preventative measures, treatments, and vaccines. In COVID-19, there have been active attempts to mine massive troves of data for these purposes (Pham et al. 2020; Vaishya et al. 2020). For instance, pharmaceutical, biomedical, and genetic data, along with scientific studies and clinical trials, are being combined and integrated for understanding how existing drugs might work in treating or preventing COVID-19.

Having accurate and complete information is also imperative in a pandemic. In this regard, conventional data sources are often inadequate. In the COVID-19 pandemic, official estimates of regional infection and fatality rates have been unreliable due to failures and delays in testing and reporting, measurement inconsistencies across organizations and places, and a high prevalence of undetected asymptomatic cases. Wastewater and sewage sensing, coupled with big data analytics, are being used in many communities to fill in these informational gaps and serve as an early warning system for COVID-19 outbreaks.

Finally, in a pandemic, it is vital to have disaggregated data, i.e., data with a high level of spatial and temporal resolution. Such data are crucial for enabling activities such as contact tracing, localized hotspot detection, and parameterization of agent-based epidemiological models. In this regard, traditional administrative records fall short, as they summarize information in an aggregated form. On the other hand, big geo-temporal data, such as that produced by “apps,” social media platforms, and mobile devices, have refined spatial and temporal granularity. Given the data are rich in information on individuals’ space-time movements and their social and spatial interaction from moment to moment, they have been an essential source of information in the pandemic.

Downsides and Dilemmas

With all that said, big data are not necessarily a magical or quick-and-easy panacea for any problem. Pandemics are no exception. First of all, there are computational and analytical challenges that

come into play, from data acquisition and filtering to analysis and modeling. Such problems are compounded by the fact that there is an information overload due to the enormous amounts of data are being produced via active and passive surveillance of people, places, and the disease itself. The quality and integrity of big data are a related matter. As with conventional sources of data, big data are far from perfect. Indeed, many big data sources used in the battle against the novel coronavirus are fraught with biases, noise, prejudices, and imperfections. For instance, social media posts, search engine queries, and Web “apps” are notoriously skewed toward particular demographics and geographies, owing to the digital divides and differences in individual preferences, needs, and desires.

The use of big data and digital analytics for managing the COVID-19 also raises various ethical and legal issues and challenges (Gasser et al. 2020; Zwitter and Gstrein 2020). One problem of grave concern in this context is privacy. Many sources of big data being used for managing the pandemic contain sensitive and personally identifiable information, which can be used to “connect the dots” about individual’s activities, preferences, and motivations. Big biometrics data, such as that produced by thermal recognition sensors, raises a similar set of concerns. While steps can be taken to mitigate privacy concerns (e.g., anonymization via the use of synthetic data), in a significant health crisis like COVID-19, there is an urgency to find solutions. Thus, the implementation of privacy protections may not be feasible or desirable, as they can hinder effective and timely public health responses.

Another set of ethical issues pertain to the use of big-enabled AI systems for decision-making in the pandemic (Leslie 2020). One problem, in particular, is that AI has the potential to produce biased and discriminatory outcomes. In general, the accuracy and fairness of AI systems hinge crucially on having precise and representative big data in the first place. Accordingly, if such data are skewed, incomplete, or inexact, AI-enabled tools and models may produce unreliable, unsafe, biased, and prejudiced outcomes and decisions (Leslie 2019). For example, facial

recognition systems – used for surveillance purposes in the pandemic – have been criticized in this regard. Further, AI learns from patterns, relationships, and dynamics associated with real-world phenomena. Hence, if there are societal gaps and disparities in the first place, then AI is likely to mimic them unless appropriate corrective actions are employed. Indeed, COVID-19 has brought to the fore various social, economic, and digital inequities, including those propelled by the pandemic itself. Accordingly, conclusions, decisions, and actions based on AI systems for the pandemic have the potential to disadvantage certain segments of the population, which has broader implications for public health, human rights, and social justice.

Looking Forward

Big data will undoubtedly play an influential role in future pandemics, which are inevitable, given our increasingly globalized society. However, as highlighted, while big data for a significant health emergency like COVID-19 brings an array of benefits and opportunities, it also comes with various downsides and dilemmas. As technology continues to accelerate and advance, and new sources of big data and analytical and computational tools surface, the upsides and downsides may look quite different as well.

Cross-References

- [Biomedical Data](#)
- [Epidemiology](#)
- [Ethical and Legal Issues](#)
- [Spatiotemporal Analytics](#)

Further Reading

- Gasser, U., Ienca, M., Scheibner, J., Sleight, J., & Vayena, E. (2020). Digital tools against COVID-19: Taxonomy, ethical challenges, and navigation aid. *The Lancet Digital Health*, 2(8), e425–e434.
- Leslie, D. (2019). Understanding artificial intelligence ethics and safety. arXiv preprint arXiv:1906.05684

- Leslie, D. (2020). Tackling COVID-19 through responsible AI innovation: Five steps in the right direction. *Harvard Data Science Review*.
- Lin, L., & Hou, Z. (2020). Combat COVID-19 with artificial intelligence and big data. *Journal of Travel Medicine*, 27(5). <https://doi.org/10.1093/jtm/taaa080>.
- Oliver, N., Letouzé, E., Sterly, H., Delataille, S., De Nadai, M., Lepri, B., et al. (2020). Mobile phone data and COVID-19: Missing an opportunity? arXiv preprint arXiv:2003.12347.
- Pham, Q. V., Nguyen, D. C., Hwang, W. J., & Pathirana, P. N. (2020). Artificial intelligence (AI) and Big Data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts. <https://doi.org/10.20944/preprints202004.0383.v1>.
- Ting, D. S. W., Carin, L., Dzau, V., & Wong, T. Y. (2020). Digital technology and COVID-19. *Nature Medicine*, 26(4), 459–461.
- Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14, 337.
- Zwitter, A., & Gstrein, O. J. (2020). Big data, privacy and COVID-19—learning from humanitarian expertise in data protection. *Journal of International Humanitarian Action*, 5(4), 1–7.

Crowdsourcing

Heather McIntosh
Mass Media, Minnesota State University,
Mankato, MN, USA

Crowdsourcing is an online participatory culture activity that brings together large, diverse sets of people and directs their energies and talents toward varied tasks designed to achieve specific goals. The concept draws on the principle that the diversity of knowledge and skills offered by a crowd exceeds the knowledge and skills offered by an elite, select few. For big data, it offers access to abilities for tasks too complex for computational analysis. Corporations, government groups, and nonprofit organizations all use crowdsourcing for multiple projects, and the crowds consist of volunteers who choose to engage tasks toward goals determined by the organizations. Though these goals may benefit the organizations more so than the crowds helping them, ideally the benefit is shared between the two. Crowdsourcing

breaks down into basic procedures, the tasks and their applications, the crowds and their makeup, and the challenges and ethical questions.

Crowdsourcing follows a general procedure. First, an organization determines the goal or the problem that requires a crowd's assistance in order to achieve or solve. Next, the organization defines the tasks needed from the crowd in order to fulfill its ambitions. After, the organization seeks the crowd's help, and the crowd engages the tasks. In selective crowdsourcing, the best solution from the crowd is chosen, while in integrative crowdsourcing, the crowd's solutions become worked into the overall project in a useful manner.

Working online is integral to the crowdsourcing process. It allows the gathering of diverse individuals who are geographically dispersed to "come together" for working on the projects. The tools the crowds need to engage the tasks also appear online. Since using an organization's own tools can prove too expensive for big data projects, organizations sometimes use social networks for recruitment and task fulfillment. The documentary project *Life in a Day*, for example, brought together video footage from people's everyday lives from around the world. When possible, people uploaded their footage to YouTube, a video-sharing platform. To address the disparities of countries without access to digital production technologies and the Internet, the project team sent cameras and memory storage cards through the mail. Other services assist with recruitment and tasks. LiveWork and Amazon Mechanical Turk are established online service marketplaces, while companies such as InnoCentive and Kaggle offer both the crowds and the tools to support an organization's project goals.

Tasks vary depending on the project's goals, and they vary in structure, interdependence, and commitment. Some tasks follow definite boundaries or procedures, while others are open-ended. Some tasks depend on other tasks for completion, while others stand alone. Some tasks require but a few seconds, while others demand more time and mental energy. More specifically, tasks might include finding and managing information, analyzing information, solving problems, and

producing content. With big data, crowds may enter, clean, and validate data. The crowds may even collect data, particularly geospatial data, which prove useful for search and rescue, land management, disaster response, and traffic management. Other tasks might include transcription of audio or visual data and tagging.

When bringing crowdsourcing to big data, the crowd offers skills that benefit through matters of judgment, contexts, and visuals – skills that exceed computational models. In terms of judgment, people can determine the relevance of items that appear within a data set, identify similarities among items, or fill in holes within the set. In terms of contexts, people can identify the situations surrounding the data and how those situations influence them. For example, a person can determine the difference between the Statue of Liberty on Ellis Island in New York and the replica on The Strip in Las Vegas. The contexts then allow determination of accuracy or ranking, such as in this case differentiating the real from the replica. People also can determine more in-depth relationships among data within a set. For example, people can better decide the accuracy of search engine terms and results matches, determine better the top search result, or even predict other people's preferences.

Properly managed crowdsourcing begins within an organization that has clear goals for its big data. These organizations can include government, corporations, and nonprofit organizations. Their goals can include improving business practices, increasing innovations, decreasing project completion times, developing issue awareness, and solving social problems. These goals frequently involve partnerships that occur across multiple entities, such as government or corporations partnering with not-for-profit initiatives.

At the federal level and managed through Massachusetts Institute of Technology's Center for Collective Intelligence, Climate CoLab brings together crowds to analyze issues related to global climate change, registering more than 14,000 members who participate in a range of contests. Within the contests, members create and refine proposals that offer climate change solutions. The proposals then are evaluated by the

community and, through voting, recommended for implementation. Contest winners presented their proposals to those who might implement them at a conference. Some contests build their initiatives on big data, such as Smart Mobility, which relies on mobile data for tracking transportation and traveling patterns in order to suggest ways for people to reduce their environmental impacts while still getting where they want to go.

Another government example comes from the city of Boston, wherein a mobile app called Street Bump tracks and maps potential potholes throughout the city in order to guide crews toward fixing them. The crowdsourcing for this initiative comes from two levels. One, the information gathered from the app helps city crews do their work more efficiently. Two, the app's first iteration reported too many false positives, leading crews to places where no potholes existed. The city worked with a crowd drawn together through InnoCentive to improve the app and its efficiency, with the top suggestions coming from a hackers group, a mathematician, and a software engineer.

Corporations also use crowdsourcing to work with their big data. AOL needed help with cataloging the content on its hundreds of thousands web pages, specifically the videos and their sources, and turned to crowdsourcing as a means to expedite and streamline the project's costs. Between 2006 and 2010, Netflix, an online streaming and mail DVD distributor, sought help with perfecting its algorithm for predicting user ratings of films. The company developed a contest with a \$1 million dollar prize, and for the contest, it offered data sets consisting of multiple million units for analysis. The goal was to beat Netflix's current algorithm by 10%, which one group achieved and took home the prize.

Not-for-profit groups also incorporate crowdsourcing as part of their initiatives. AARP Foundation, which works on behalf of older Americans, used crowdsourcing to tackle such issues as eliminating food insecurity and food deserts (areas where people do not have convenient or close access to grocery stores). Humanitarian Tracker crowdsources data from people "on the ground" about issues such as disease, human rights violations, and rape. Focusing particularly

on Syria, Humanitarian Tracker aggregates these data into maps that show the impacts of systematic killings, civilian targeting, and other human tolls.

Not all crowdsourcing and big data projects originate within these organizations. For example, Galaxy Zoo demonstrates the expanses of both big data and crowds. The project asked people to classify a data set of one million galaxies into three categories: elliptical, merger, and spiral. By the project's completion, 150,000 people had contributed 50 million classifications. The data feature multiple independent classifications as well, adding reliability. The largest crowdsourcing project involved searching satellite images for wreckage from Malaysia Airlines flight MH370, which went missing in March 2014. Millions of people searched for signs among the images made available by Colorado-based Digital Globe. The amount of crowdsourcing traffic even crashed websites.

Not all big data crowdsourced projects succeed, however. One example is the Google Flu tracker. The tracker included a map to show the disease's spread throughout the season. It was later revealed that the tracker overestimated the expanse of the flu spreading, predicting twice as much as actually occurred.

In addition to their potentially not succeeding, another drawback to these projects is their overall management, which tends to be time-consuming and difficult. Several companies attempt to fulfill this role. InnoCentive and Kaggle use crowds to tackle challenges brought to them by industries, government, and nonprofit organizations. Kaggle in particular offers almost 150,000 data scientists – statisticians – to help companies develop more efficient predictive models, such as deciding the best order in which to show hotel rooms for a travel app or guessing which customers would leave an insurance company within a year. Both InnoCentive and Kaggle run their crowdsourcing activities as contests or competitions as these are often tasks that require a higher time and mental commitment than others.

Crowds bring wisdom to crowdsourced tasks on big data through their diversity of skills and knowledge. Determining the makeup of that crowd proves more challenging, but one study of

Mechanical Turk offers some interesting findings. It found that US females outnumber males by 2 to 1 and that many of the workers hold bachelor's and even master's degrees. Most live in small households of two or fewer people, and most use the crowdsourcing work to supplement their household incomes as opposed to being the primary source of income.

Crowd members choose the projects on which they want to work, and multiple factors contribute to their motivations for joining a project and staying with it. For some working on projects that offer no further incentive to participate, the project needs to align with their interests and experience so that they feel they can make a contribution. Others enjoy connecting with other people, engaging in problem-solving activities, seeking something new, learning more about the data at hand, or even developing a new skill. Some projects offer incentives such as prize money or top-contributor status. For some entertainment motivates them to participate in that the tasks offer a diversion. For others, though, working on crowdsourced projects might be addiction as well.

While crowdsourcing offers multiple benefits for the processing of big data, it also draws some criticism. A primary critique centers on the notion of labor, wherein the crowd contributes knowledge and skills for little-to-no pay, while the organization behind the data stands to gain much more financially. Some crowdsourcing sites offer low cash incentives for the crowd participants, and in doing so, they sidestep labor laws requiring minimum wage and other worker benefits. Opponents of this view cite that the labor involved frequently requires menial tasks and that the labor faces no obligation in completing the assigned tasks. They also cite that crowd participants engage the tasks because they enjoy doing so.

Ethical concerns come back to the types of crowdsourced big data projects and the intentions behind them, such as information gathering, surveillance, and information manipulation. With information manipulation, for example, crowd participants might create fake product reviews and ratings for various web sites, or they might crack anti-spam devices such as CAPTCHAs (Completely Automated Public Turing test to tell

Computers and Humans Apart). Other activities involve risks and possible violations of other individuals, such as gathering large amounts of personal data for sale. Overall, the crowd participants remain unaware that they are engaging in unethical activities.

Cross-References

- [Cell Phone Data](#)
- [Netflix](#)
- [Predictive Analytics](#)

Further Reading

- Brabham, D. C. (2013). *Crowdsourcing*. Cambridge, MA: MIT Press.
- Howe, J. (2009). *Crowdsourcing: why the power of the crowd is driving the future of business*. New York: Crown.
- Nakatsu, R. T., Grossman, E. B., & Charalambos, L. I. (2014). A taxonomy of crowdsourcing based on task complexity. *Journal of Information Science*, 40(6), 823–834.
- Shirky, C. (2009). *Here comes everybody: the power of organizing without organizations*. New York: Penguin.
- Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor.

Cultural Analytics

Tobias Blanke
Department of Digital Humanities, King's
College London, London, UK

Definition

Cultural analytics was originally introduced by Lev Manovich in 2007 in order to describe the use of “computational and visualization methods for the analysis of massive cultural data sets and flows” and “to question our basic cultural concepts and methods” (Software Studies Initiative 2014) Manovich was then especially concerned with “the exploration of large cultural data sets by

means of interactive and intuitive visual analysis techniques” (Yamaoka et al. 2011) and massive multimedia data sets (Manovich 2009).

In 2016, Manovich further elaborated that cultural analytics brings together the disciplines of digital humanities and social computing. “[W]e are interested in combining both in the study of cultures – focusing on the particular, interpretation, and the past from the humanities, while centering on the general, formal models, and predicting the future from the sciences” (Manovich 2016). Cultural analytics works with “historical artifacts” as well as “the study of society using social media and social phenomena specific to social networks.” It can thus be summarized as any kind of advanced computational technique to understand digital cultural expressions, as long as these reach a certain size.

Big Cultural Data

Big data is not limited to the sciences and large-scale enterprises. With more than seven billion people worldwide and counting, vast amounts of data are produced in social and cultural interactions. At the same time, we can look back onto several thousand years of human history that have delivered huge numbers of cultural records. Large-scale digitization efforts have recently begun to create digital surrogates for these records that are freely available online. In the USA, the HathiTrust published research data extracted from over 4,800,000 volumes of digitized books (containing 1.8 billion pages) – including parts of the Google Books corpus and the Internet Archive. The European Union continues to be committed to digitize and present its cultural heritage online. At the time of writing, its cultural heritage aggregator Europeana has made available over 60 m digital objects (Heath 2014).

Nature covered the topic of big cultural data already in 2010 (Hand 2011) and compared typical data sets used in cultural research with those in the sciences that can be considered big. While data sets from the Large Hadron Collider are still by far the largest around, cultural data sets can easily compare to other examples of big sciences.

The Sloan Digital Sky Survey, for instance, had brought together about 100 TB of astronomical observations by the end of 2010. This is big data, but not as big as some cultural heritage data sets. The Holocaust Survivor Testimonials’ Collections by the Shoah Foundation contained 200 TB of data in 2010. The Google Books corpus had hundreds of millions of books. Another typical digitization project, the Taiwanese TELDAP archive of Chinese and Taiwanese heritage objects, had over 250 TB of digitized content in 2011 (Digital Taiwan 2011).

Book corpora like the Google Books project or historical testimonials such as the Holocaust Survivor Testimonials are the primary type of data associated with digital culture. Quantitative work with these has been popularized in the work on a “Quantitative Analysis of Culture” by Michel et al. (2011), summarizing the big trends of human thoughts with the Google Ngram corpus and thus moving to corpora that cannot be read by humans alone, because they are too large. From a scholarly point of view, Franco Moretti (2000, 2005) has pioneered quantitative methods to study literature and advocates “distant reading” and “a unified theory of plot and style” (Schulz 2011). The methods Moretti uses such as social network analysis have been employed by social scientists for a long time but hardly in the study of culture. An exception is the work by Schich et al. (2014) to develop a network framework of cultural history of the lives of over 100,000 historical individuals. Other examples of new quantitative methods for the large-scale study of digital culture include genre detection in literature. Underwood et al. (2013) demonstrated how genres can be identified in the HathiTrust Digital Library corpus in order to “trace the changing proportions of first- and third-person narration,” while computational stylistics is able to discover differences in the particular choices made by individuals and groups using languages (Eder 2016). This has now become a fast-developing new field, brought together in the recently launched *Journal of Cultural Analytics* (<http://culturalanalytics.org/>).

Next to such digitized cultural sources, cultural analytics, however, also works with the new digital materials that can be used to capture

contemporary culture. At its eighth birthday in 2013, the YouTube video-sharing site proudly announced that “more than 100 hours of video are uploaded to YouTube every minute” (YouTube 2013), while the site has been visited by over one billion users, many of which have added substantial cultural artifacts to the videos such as annotations or comments. Facebook adds 350 million photos to its site every day. All these are also cultural artifacts and are already now subject to research on contemporary culture. It was against the background of this new cultural material that Manovich formulated his idea of cultural analytics, which is interested not just in great historical individuals but in “everything created by everybody” (Manovich 2016).

Commercial Implications

Because cultural analytics is interested in everything created by everybody, it rushes to work with new cultural materials in social media. It does so not solely because of the new kinds of research in digital culture but also because of new economic value from digital culture. Social media business models are often based on working with cultural analytics using social computing as well as digital humanities. The almost real-time gathering of opinions (Pang and Lee 2008) to read the state of mind of organizations, consumers, politicians, and other opinion makers continues to excite businesses around the world. Twitter, YouTube, etc. now appear to be the “echo chamber of people’s opinions” (Van Dijck and Poell 2013, p. 9). Companies, policy researchers, and many others have always depended on being able to track what people believe and think. Social media cultural artifacts can signal the performance of stocks by offering insights on the emotional state of those involved with companies. In this way, they allow for measuring the culture of a company, of groups, and of individuals and have led to a new research area called “social sensing” (Helbing 2011).

Cultural analytics has shown, for instance, that online political cultures are not that different from the real world (Rainie 2014). Twitter’s political

culture is polarized around strong political ties and works often as an echo chamber for one’s already formed political opinions. Twitter users of the same political pervasion cluster together in fragmented groups and do not take input from the outside. This kind of cultural analytics is critical toward analyzing the democratizing effect of social media. Just because there is communication happening, this communication does not necessarily lead to political exchange across traditional boundaries.

Big money is currently flowing into building cultural analytics engines that help understand users’ preferences, social media likes, etc. This money is effectively spent on analyzing and monetizing our digital culture, which has become the currency with which we pay for online services. John Naughton has pointed out that we “pay” for all those free online services nowadays in a “different currency, namely your personal data” (Naughton 2013). Because this personal data is never just personal but involves others, he could have also said that we pay with our digital culture. Cultural analytics algorithms decide how we should fill our shopping basket, which political groups we should join on Facebook, etc. However, the data used in these commercial cultural analytics and produced by it is seldom open to those who produce it (Pybus et al. 2015). The companies and organizations like Google, who collect this data, also own the underlying cultural data and its life cycle and therefore our own construction of our digital identity.

The cultural analytics done by the large Internet intermediaries such as Google and Facebook means that cultural expressions are quantified and analyzed for their commercial value – in order to market and sell more products. However, not everything can be quantified, and a critical approach to cultural analytics also needs to understand what is lost in such quantification and what its limits are. We have already discussed, e.g., how Twitter’s organization around counting retweets and followers makes its users encounter mainly more of the same in political interactions. Understanding the limits of cultural analytics will be a major role for the study of digital culture in the future that will also need to identify how

opposition against such commercial cultural analytics practices can be formulated and practiced.

Critique of Cultural Analytics

The study of digital culture needs to remain critical to what is possible when algorithms are used to understand culture. It has already begun to do so in its critical analysis of the emerging tools and methods of cultural analytics. A good example is the reaction to the already discussed Google Ngram viewer ([http:// books.google.com/ngrams](http://books.google.com/ngrams)), which enables public access to the cultural analytics of millions of words in books. Erez Aiden and Jean-Baptiste Michel (pioneers of the Google Ngram Viewer) go as far as to promise a “new lens on human culture” and a transformation of the scientific disciplines concerned with observing it. The Ngram Viewer’s “consequences will transform how we look at ourselves (...). Big data is going to change the humanities, transform the social sciences, and renegotiate the relationship between the world of commerce and the ivory tower” (Aiden and Michel 2013).

This kind of enthusiasm is at least a little surprising, because it is not easy to find the exciting new research that the Ngram viewer has made possible. Pechenick et al. (2015) have demonstrated the “strong limits to inferences of socio-cultural and linguistic evolution” the Ngram viewer allows because of severe shortages in the underlying data. Researchers also complain about “trivial” results research the Ngram viewer delivers (Kirsch 2014). No cultural historian needs the Ngram viewer to understand that religion is in retreat in the nineteenth century. This does not mean that the Ngram viewer cannot produce new kinds of evidence or even new insights, but this needs to be carefully examined and involve the critical interpretation of primary and secondary sources using traditional approaches next to cultural analytics.

While the Ngram viewer is an interesting tool for research and education, it is exaggerated to claim that it is already changing cultural research in a significant way. Against Moretti and the Google Ngram efforts, for some researchers of culture,

we should rather be interested in what “big data will never explain,” as Leon Wieseltier has put it:

In the riot of words and numbers in which we live so smartly and so articulately, in the comprehensively quantified existence in which we presume to believe that eventually we will know everything, in the expanding universe of prediction in which hope and longing will come to seem obsolete and merely ignorant, we are renouncing some of the primary human experiences. (Wieseltier 2013)

Leon Wieseltier has emerged as one of the strongest opponents of cultural analytics: “As even some partisans of big data have noted, the massive identification of regularities and irregularities can speak to ‘what’ but not to ‘why’: they cannot recognize causes and reasons, which are essential elements of humanistic research” (Wieseltier 2013). Responding to such criticisms, for Manovich cultural analytics, should therefore not just focus on the large trends but on individuals, true to its humanistic foundations. “[W]e may combine the concern of social science, and sciences in general, with the general and the regular, and the concern of humanities with individual and particular” (Manovich 2016).

Conclusions

While the criticisms by Wieseltier and others should be taken serious and emphasize important limitations and the dangers of a wrong kind of cultural analytics, the study of culture also needs to acknowledge that many of the digital practices associated with cultural analytics show promise. More than 10 years ago, Google and others revolutionized information access, while Facebook has allowed for new kinds of cultural connections since the early 2000s. This kind of efforts has made us understand that there are many more books available than any single person in the world can read in a lifetime and that computers can help us with this information overload and stay on top of the analysis. It is the foremost task of cultural analytics to understand better how we can use new digital tools and techniques in cultural research, which includes understanding the boundaries and what they cannot explain.

Further Reading

- Aiden, E., & Michel, J.-B. (2013). *Uncharted: Big data as a lens on human culture*. New York: Penguin.
- Digital Taiwan. (2011). NDAP international conference. http://culture.teldap.tw/culture/index.php?option=com_content&view=article&id=23:ndap-international-conference-&catid=1:events&Itemid=215. Accessed 2 July 2016.
- Eder, M. (2016). Rolling stylometry. *Digital Scholarship Humanities*, 31(3), 457–469.
- Hand, E. (2011). Culturomics: Word play. *Nature*, 474 (7352), 436–440.
- Heath, P. (2014). Europe's cultural heritage online. <https://ephthinktank.eu/2014/04/09/europes-cultural-heritage-online/>. Accessed 2 July 2016.
- Helbing, D. (2011). FuturICT—A knowledge accelerator to explore and manage our future in a strongly connected world. *arXiv preprint arXiv:1108.6131*.
- Kirsch, A. (2014). Technology is taking over English departments. <https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch>. Accessed 2 July 2016.
- Manovich, L. (2009). Cultural analytics: Visualizing cultural patterns in the era of more media. *Domus*, 923.
- Manovich, L. (2016). The science of culture? Social computing, digital humanities and cultural analytics. *Cultural Analytics*, 1(1).
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J., & Orwant, J. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.
- Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 1, 54–68.
- Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.
- Naughton, J. (2013). To the internet giants, you're not a customer. You're just another user, The Guardian. <http://www.theguardian.com/technology/2013/jun/09/internet-giants-just-another-customer>. Accessed 2 July 2016.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.
- Peckenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS One*, 10(10), e0137041.
- Pybus, J., Coté, M., Blanke, T. (2015). Hacking the social life of big data. *Big Data & Society*, 2(2).
- Rainie, L. (2014). The six types of twitter conversations. Pew Research Center. <http://www.pewresearch.org/fact-tank/2014/02/20/the-six-types-of-twitter-conversations/>. Accessed 2 July 2016.
- Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. *Science*, 345(6196), 558–562.
- Schulz, K. (2011). What is distant reading? http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?_r=0. Accessed 2 July 2016.
- Software Studies Initiative. (2014). Cultural analytics. <http://lab.softwarestudies.com/p/cultural-analytics.html>. Accessed 2 July 2016.
- Underwood, T., Black, M.L., Auvil, L., & Capitanu, B. (2013). Mapping mutable genres in structurally complex volumes. 2013 I.E. International Conference on Big Data. Washington, DC: IEEE.
- Van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, 1(1), 2–14.
- Wieseltier, L. (2013). What big data will never explain, New Republic. <http://www.newrepublic.com/article/112734/what-big-data-will-never-explain>. Accessed 2 July 2016.
- Yamaoka, S., Manovich, L., Douglass, J., & Kuester, F. (2011). Cultural analytics in large-scale visualization environments. *Computer*, 44(12), 39–48.
- YouTube. (2013). Here's to eight great years. From <http://youtube-global.blogspot.co.uk/2013/05/heres-to-eight-great-years.html>. Accessed 2 July 2016.

Curriculum, Higher Education, and Social Sciences

Stephen T. Schroth
Department of Early Childhood Education,
Towson University, Baltimore, MD, USA

Big data, which has revolutionized many practices in business, government, healthcare, and other fields, promises to radically change the curriculum offered in many of the social sciences. Big data involves the capture, collection, storage, collation, search, sharing, analysis, and visualization of enormous data sets so that this information may be used to spot trends, to prevent problems, and to proactively engage in activities that make success more likely. The social sciences, which includes fields as disparate as anthropology, economics, education, political science, psychology, and sociology, is a disparate area, and the tools of big data are being embraced differently within each. The economic demands of setting up systems that permit the use of big data in higher education have also hindered some efforts to use these processes, as these institutions often lack the infrastructure necessary to proceed with such efforts. Opponents of the trend toward using big

data tools for social science analyses often stress that while these tools may provide helpful for certain analyses, it is also crucial for students to receive training in more traditional methods. As equipment and training concerns are overcome, however, the use of big data social sciences departments at colleges and universities seems likely to increase.

Background

A variety of organizations, including government agencies, businesses, colleges, universities, schools, hospitals, research centers, and others, collect data regarding their operations, clients, students, patients, and findings. Disciplines within the social sciences, which are focused upon society and the relationships among individuals within a society, often use such data to inform studies related to these. Such a volume of data has been generated, however, that many social scientists have found it impossible to use this in their work in a meaningful manner. The emergence of computers and other electronic forms of data storage has resulted in more data than ever before being collected, especially during the last two decades of the twentieth century. This data was generally stored in separate databases. This worked to make data from different sources inaccessible to most social science users. As a result, much of the information that could potentially be obtained from such sources was not used.

Over the past decade and a half, many businesses became increasingly interested in making use of data they had but did not use regarding customers, processes, sales, and other matters. Big data became seen as a way of organizing and using the numerous sources of information in ways that could benefit organizations and individuals. Infonomics, the study of how information could be used for economic gain, grew in importance as companies and organizations worked to make better use of the information they possessed, with the end goal being to use it in ways that increased profitability. A variety of consulting firms and other organizations began working

with large corporations and organizations in an effort to accomplish this. They defined *big data* as consisting of three “v”s, volume, variety, and velocity.

Volume, as used in this context, refers to the increase in data volume caused by technological innovation. This includes transaction-based data that has been gathered by corporations and organizations over time but also includes unstructured data that derives from social media and other sources as well as increasing amounts of sensor and machine-to-machine data. For years, excessive data volume was a storage issue, as the cost of keeping much of this information was prohibitive. As storage costs have decreased, however, cost has diminished as a concern. Today, how best to determine relevance within large volumes of data and how best to analyze data to create value have emerged as the primary issues facing those wishing to use it.

Velocity refers to the amount of data streaming in at great speed raises the issue of how best to deal with this in an appropriate way. Technological developments, such as sensors and smart meters, and client and patient needs emphasize the necessity of overseeing and handling inundations of data in near real time. Responding to data velocity in a timely manner represents an ongoing struggle for most corporations and other organizations. *Variety* in the types of formats in which data today comes to organizations presents a problem for many. Data today includes that in structured numeric forms which is stored in traditional databases but has grown to include information created from business applications, e-mails, text documents, audio, video, financial transactions, and a host of others. Many corporations and organizations struggle with governing, managing, and merging different forms of data.

Some have added two additional criteria to these: variability and complexity. *Variability* concerns the potential inconsistency that data can demonstrate at times, which can be problematic for those who analyze the data. Variability can hamper the process of managing and handling the data. *Complexity* refers the intricate process that data management involves, in particular when

large volumes of data come from multiple and disparate sources. For analysts and other users to fully understand the information that is contained in these data, they must be must first be connected, correlated, and linked in a way that helps users make sense of them.

Big Data Comes to the Social Sciences

Colleges, universities, and other research centers have tracked the efforts of the business world to use big data in a way that helped to shape organizational decisions and increase profitability. Many working in the social sciences were intrigued by this process, as they saw it as a useful tool that could be used in their own research. The typical program in these areas, however, did not provide students, be they at the undergraduate or graduate level, the training necessary to engage in big data research projects. As a result, many programs in the social sciences have altered their curriculum in an effort to assure that researchers will be able to carry out such work. For many programs across the social sciences that have pursued curricular changes that will enable students to engage in big data research, these changes have resulted in more coursework in statistics, networking, programming, analytics, database management, and other related areas. As many programs already required a substantial number of courses in other areas, the drive toward big data competency has required many departments to reexamine the work required of their students.

This move toward more coursework that supports big data has not been without its critics. Some have suggested that changes in curricular offerings have come at a high cost, with students now being able to perform certain operations involved with handling data but unable to competently perform other tasks, such as establishing a representative sample or composing a valid survey. These critics also suggest that while big data analysis has been praised for offering tremendous promise, in truth the analysis performed is shallow, especially when compared to that done with smaller data sets. Indeed, representative sampling

would negate the need for, and expense of, many big data projects. Such critics suggest that increased emphasis in the curriculum should focus on finding quality, rather than big, data sources and that efforts to train students to load, transform, and extract data is sublimating other more important skills.

Despite these criticisms, changes to the social sciences curriculum are occurring at many institutions. Many programs now require students to engage in work that examines practices and paradigms of data science, which would provide students with a grounding in the core concepts of data science, analytics, and data management. Work in algorithms and modeling, which provide proficiency in basic statistics, classification, cluster analysis, data mining, decision trees, experimental design, forecasting, linear algebra, linear and logistic regression, market basket analysis, predictive modeling, sampling, text analytics, summarization, time series analysis, and unsupervised learning constrained optimization, is also an area of emphasis in many programs. Students require exposure to tools and platforms, which provides proficiency in modeling, development and visualization tools to be used on big data projects, as well as knowledge about the platforms used for execution, governance, integration, and storage of big data. Finally, work with applications and outcomes, which emphasize the primary applications of data science to one's field, and how it interacts with disciplinary issues and concerns have been emphasized by many programs.

Some programs have embraced big data tools but suggested that not every student needs mastery of them. Instead, these programs have suggested that big data has emerged as a field of its own and that certain students should be trained in these skills so that they can work with others within the discipline to provide support for those projects that require big data analysis. This approach offers more incremental changes to the social science curricular offerings, as it would require fewer changes for most students yet still enable departments to produce scholars who are equipped to engage in research projects involving big data.

Cross-References

- [Big Data Quality](#)
- [Correlation Versus Causation](#)
- [Curriculum, Higher Education, and Social Sciences](#)
- [Curriculum, Higher Education, Humanities](#)
- [Education](#)

Further Reading

- Foreman, J. W. (2013). *Data smart: Using data science to transform information into insight*. Hoboken: Wiley.
- Lane, J. E., & Zimpher, N. L. (2014). *Building a smarter university: Big data, innovation, and analytics*. Albany: The State University of New York Press.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data*. New York: Mariner Books.
- Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Hoboken: Wiley.

Curriculum, Higher Education, Humanities

Ulrich Tiedau

Centre for Digital Humanities, University College London, London, UK

Introduction

As a relatively new field, there is no generally accepted standard or reference curriculum for big data in the Humanities yet. This entry highlights some of the main strands and common themes that seem to be emerging and provides pointers to further resources.

While quantitative methods, often borrowed from the Social Sciences, especially statistics and content analysis, and corresponding software packages (e.g., SAS, SPSS, STATA), have been part of the curriculum of more social science-orientated subjects such as social and economic history, for decades, introductions to big data analysis in literary studies, English, and modern

foreign languages programs, let alone in other Humanities subjects, have been far and between until fairly recently. A notable exception in this respect are Classical Studies, in which the availability of a limited, well-defined, and comparatively small corpus of ancient texts (in a pre-WWW age fitting on one or a couple of CD-ROMs, e.g., the Perseus Digital Library) has lent itself to some form of corpus analytics; and historical subjects with similarly well-described limited corpora, e.g., medieval studies, plus of course corpus linguistics itself, in which computational methods have long figured prominently.

Having said that, the size of corpora available since the introduction, success and exponential growth of the World Wide Web, e.g., Google Books with its 25 million and growing number of digitized books (2015), outstrips the size of previously available corpora by several orders of magnitude so that “big data” also here can still be seen as a fairly recent development.

Types of Courses

Digital approaches to the Humanities are taught in departments across the whole spectrum of the Humanities, with subjects such as English and History leading in terms of numbers; followed by Media Studies and Library, Archive, and Information Studies; dedicated Digital Humanities programs and interdisciplinary programs, e.g., liberal arts and sciences degrees, not to forget, on the boundaries of engineering and the Humanities, Computer Science (Spiro 2011). Especially in the USA, English departments seem to have taken the lead in the Digital Humanities (Kirschenbaum 2010), whereas History also has a long and distinctive tradition in Digital History (Cohen and Rosenzweig 2005), leading to recent discussions whether or not Digital History can be considered a separate development from Digital Humanities or an integral part of it (e.g., Robertson 2014; McGinnis 2014). There are also general methodological courses aimed at students

of all Humanities subjects at a great number of institutions.

While most of these courses are self-contained and usually optional modules in their respective curricula, dedicated Digital Humanities programs provide systematic introduction. Since the mid-2000s, these specialist degree courses, of which some focus more on the cultural side of Digital Humanities, whereas others have a pronounced emphasis on technology, are rapidly emerging at higher education institutions all over the world; Melissa Terras's visualization of the spread of Digital Humanities (2012) for example counts 114 physical DH centers in 24 countries in 2012. Established postgraduate degree programs exist at places like Kings' College London, Loyola University Chicago, the University of Alberta, University College London, and University College Cork, Ireland (Gold 2011; cf. Centernet.org for a full list).

Common Themes

In a first analysis of 134 curricula of DH courses, Lisa Spiro (2011) observes three common general characteristics: firstly, that most courses make a connection between theory and practice by explicitly or implicitly employing a project-based learning (PBL) approach, requiring students to get involved in hands-on or practice learning by building digital artifacts. This reflects the double nature of Digital Humanities, that it is as much about *building* websites, databases, or demonstrators as about *analyzing* and *interpreting* them, a theory-practice dichotomy that only at first sight seems to be new as Kathleen Fitzpatrick (2011) has pointed out, as it exists in other areas of the Humanities as well, e.g., in the separation of Creative Arts from Art History, or of Creative Writing from Literary Analysis. DH in this respect is overarching the divide.

Secondly, in line with DH research culture, DH courses not only teach applying tools, methods, and technology but also group work and

collaboration, an innovation just as transformative to traditional Humanities research culture with its "lone scholar ideal" as the use of computational methods. Often this aspect of the curriculum also includes project management, thus training key skills that are also relevant in other contexts (cf. Mahony et al. 2012). And thirdly, again in line with DH culture, open practice and digital scholarship figure prominently, frequently requiring students to practice in the open, keeping learning journals, using blogs, wikis, social media like Twitter, etc.

In terms of technologies taught, most digital courses in the Humanities, traditionally predominantly text-based subjects, unsurprisingly focus on text analysis and text-encoding. XML and TEI are the most frequently taught technologies here and plenty of free teaching resources are available (e.g., TEI by Example, <http://www.teiByExample.org>; DHOER, <http://www.ucl.ac.uk/dhoer>). As tools for analysis and visualization, Google's n-gram viewer (<http://books.google.com/ngrams>) and Voyant Tools (<http://www.voyant-tools.org>) are popular, both due to their ease of use and their not requiring knowledge of any coding. Besides text encoding and analytics, processing of sound and still and moving images, databases and networks, simulations and games, maps and geospatial information systems (GIS), and data visualization are also part of course syllabi (Spiro 2011).

A major debate seems to be about whether or not the curriculum needs to include coding, in other words whether you can pursue Digital Humanities without being able to program. While there are certainly arguments in favor of coding, many modern tools, some of them specifically designed with a teaching and learning aspect in mind (e.g., Omeka for the presentation of digital artifacts (<http://www.omeka.org>), Neatline for geotemporal visualization of Humanities data (<http://www.neatline.org>)) do not require any coding skills at all. Neither do the frequently used Google n-gram viewer (<http://books.google.com/ngrams>) for basic and Voyant Tools (<http://www.voyant-tools.org>) for more

advanced text-mining and textual analytics. On the other hand, straight-forward text-books and online introductions to computer-assisted text analysis using the programming and scripting languages R, Python, PHP (MySQL), SPARQL (RDF), and others are available, specifically directed at Humanities scholars wishing to acquire the necessary coding skills, whether in the classroom or in self-study (e.g., Jockers 2013; and in Hirsch 2012).

Further Resources

The largest collection of Digital Humanities course syllabi is openly available via Lisa Spiro's *Digital Humanities Education Zotero Group* (https://www.zotero.org/groups/digital_humanities_education), for selections of particularly relevant and recent ones also see Gold 2012 and Hancher 2014. An important source and discussion platform for pedagogical and curricular information is the website of the Humanities, Arts, Science and Technology Alliance and Collaboratory (HASTAC), a virtual organization of more than 12,000 individuals and institutions dedicated to innovative new modes of learning and research in higher education (<http://www.hastac.org>).

Cross-References

- [Big Humanities Project](#)
- [Humanities \(Digital Humanities\)](#)

Further Reading

- Cohen, D., & Rosenzweig, R. (2005). *Digital history: A guide to gathering, preserving and presenting the past on the web*. Philadelphia: University of Pennsylvania Press.
- Fitzpatrick, K. The humanities done digitally. *The chronicle of higher education*. 8 May 2011. <http://chronicle.com/article/The-Humanities-Done-Digitally/127382/>. Accessed Aug 2014.
- Gold, M. K. Digital humanities syllabi (6 June 2011). <http://cunydh.commons.gc.cuny.edu/2011/06/06/digital-humanities-syllabi/>. Accessed Aug 2014.

- Gold, M. (Ed.). (2012). *Debates in the Digital Humanities*. Minneapolis: Minnesota University Press.
- Hancher, M. Recent digital humanities syllabi (18 January 2014). <http://blog.lib.umn.edu/mh/dh2/2014/01/recent-digital-humanities-syllabi.html>. Accessed Aug 2014.
- Hirsch, B. D. (Ed.). (2012). *Digital humanities pedagogy: Practices, principles and politics*. Cambridge: Open Book Publishers.
- Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.
- Jockers, M. L. (2014). *Text analysis with R for students of literature*. Heidelberg/New York: Springer.
- Kirschenbaum, M. G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin*, (150), 1–7. <https://doi.org/10.1632/ade.150.55>.
- Mahony, S., Tiedau, U., & Sirmons, I. (2012). Open access and online teaching materials in digital humanities. In C. Warwick, M. Terras, & J. Nyhan (Eds.), *Digital humanities in practice* (pp. 168–191). London: Facet.
- Paul McGinnis, P. (2014). DH vs. DH, and Moretti's war. <http://majining.com/?p=417>. Accessed Aug 2014.
- Robertson, S. The differences between digital history and digital humanities. <http://drstephenrobertson.com/blog-post/the-differences-between-digital-history-and-digital-humanities/>. Accessed Aug 2014.
- Spiro, L. (2011). Knowing and doing: Understanding the digital humanities curriculum. June 2011. <http://digitalscholarship.files.wordpress.com/2011/06/spirodheducationpresentation2011-4.pdf>. Accessed Aug 2014.
- Spiro, L. (2014). Shaping (digital) scholars: Design principles for digital pedagogy. <https://digitalscholarship.files.wordpress.com/2014/08/spirodigitalpedagogyuts2014.pdf>.
- Spiro, L. Digital Humanities Education Zotero Group. https://www.zotero.org/groups/digital_humanities_education. Accessed Aug 2014.

Cyber Espionage

David Freet¹ and Rajeev Agrawal²

¹Eastern Kentucky University, Southern Illinois University, Edwardsville, IL, USA

²Information Technology Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, USA

Introduction

Cyber espionage or cyber spying is the act of obtaining personal, sensitive, or proprietary

information from individuals without their knowledge or consent. In an increasingly transparent and technological society, the ability to control the private information an individual reveals on the Internet and the ability of others to access that information are a growing concern. This includes storage and retrieval of e-mail by third parties, social media, search engines, data mining, GPS tracking, the explosion of smartphone usage, and many other technology considerations. In the age of big data, there is growing concern for privacy issues surrounding the storage and misuse of personal data and non-consensual mining of private information by companies, criminals, and governments.

Concerning the growing threat of cyber espionage in the big data world, Sigholm and Bang write that unlike traditional crimes, companies cannot call the police and expect them to pursue cyber criminals. Affected organizations play a leading role in each and every investigation because it is their systems and data that are being stolen or leveraged. The fight against cybercrime must be waged on a collective basis, regardless of whether the criminal is a rogue hacker or a nation-state (Sigholm and Bang 2013).

In 1968, the US government passed the Omnibus Crime Control and Safe Streets Act, which included a wiretapping law that became commonly known as the Wiretap Act (Burgunder 2011, p. 462). This law made it illegal for any person to willfully use an electronic or mechanical device to intercept an oral communication unless prior consent was given or if the interception occurred during the ordinary course of business. In 1986, Congress passed the Electronic Communications Privacy Act (ECPA) which amended the original Wiretap Act and also introduced the Stored Communications Act (SCA) which primarily prevents outsiders from hacking into facilities that are used to store electronic communications. These pieces of legislation form the cornerstone for defining protections against cyber espionage in the age of big data and social media.

In contrast to US privacy laws, the European Union (EU) has adopted significant legislation governing the collection and processing of personal information. This ensures that personal

data is processed within acceptable privacy limits and that informed consent is present. Compared to the EU, the USA has relatively few laws that enforce information privacy. The USA has mostly relied on industry guidelines and practices to ensure privacy of personal information, and the most significant feature of EU regulations in relation to the USA has been the prohibition of the transfer of personal data to countries outside the EU that do not guarantee an adequate level of protection (Burgunder 2011, p. 478). EU law affirms that the collection, storage, or disclosure of information relating to private life interferes with the right to private life and therefore requires justification.

Ironically, as we become a more technologically dependent society with increased public surveillance, data mining, transparency of private information, and social media, we come to expect less privacy and are consequently entitled to less of it (Turley 2011). This leads to the difficult question of how much privacy we as individuals can “reasonably” expect. Recently, the *Jones v. United States* case challenged the existing privacy expectations over police surveillance using a GPS that monitored the suspect’s location. As the normalcy of warrantless surveillance increases, our expectations fall, allowing this type of surveillance to become more “common.” This results in a move toward limitless police powers. These declining expectations are at the heart of the Obama administration’s argument in this case, where it affirms that the government is free to track citizens without warrants because citizens “expect to be monitored” (Turley 2011).

Vulnerable Technologies

Smartphones have become a common fixture of daily life, and the enormous amount of personal data stored on these devices has led to unforeseen difficulties with the interpretation of laws meant to protect privacy. Current legislation has actually focused on defining a smartphone as an extension to an individual’s home for the sake of protecting sensitive information. In 2009, the Supreme Court of Ohio issued the most clear-cut case which held

that the search of a smartphone incident to an arrest is unreasonable under the Fourth Amendment. The court held in *State v. Smith* that because a smartphone allows for high-speed Internet access and is capable of storing “tremendous amounts of private data,” it is unlike other containers for the purposes of Fourth Amendment analysis. Because of this large amount of personal information, its user has a “high expectation” of privacy. In short, the state may confiscate the phone in order to collect and preserve evidence but must then obtain a warrant before intruding into the phone’s contents (Swingle 2012, p. 37). Smartphones have evolved into intimate collections of our most personal data and no longer just lists of phone numbers, the type of data that has traditionally been kept in the privacy of our homes and not in our pockets.

The phenomenon of social media has raised a host of security and privacy issues that never existed before. The vast amount of personal information displayed and stored on sites such as Facebook, Snapchat, MySpace, and Google make it possible to piece together a composite picture of users in a way never before possible. Social networking sites contain various levels of privacy offered to users. Sites such as Facebook encourage users to provide real names and other personal information in order to develop a profile that is then available to the public. Many online dating sites allow people to remain anonymous and in more control of their personal data. This voluntarily divulgence of so much personal information plays into the debate over what kind of privacy we can “reasonably expect.”

In 2003, an individual named Kathleen Romano fell off an allegedly defective desk chair while at work. Romano claimed she sustained serious permanent injuries involving multiple surgeries and sued Steelcase Inc., the manufacturer of the chair. Steelcase refuted the suit saying Romano’s claims of being confined to her house and bed were unsubstantiated based on public postings on her Facebook and MySpace profiles which showed her engaged in travel and other rigorous physical activities (Walder 2010). When Steelcase

attempted to procure these pictures as evidence, Romano opposed the motion claiming she “possessed a reasonable expectation of privacy in her home computer” (Walder 2010). Facebook also opposed releasing Romano’s profile information without her consent because it violated the federal Stored Communications Act. Acting Justice Jeffrey Arlen Spinner of New York’s Suffolk County Supreme Court rejected Romano’s argument that the release of information would violate her Fourth Amendment right to privacy. Spinner wrote “When Plaintiff created her Facebook and MySpace accounts she consented to the fact that her personal information would be shared with others, notwithstanding her privacy settings. Indeed, that is the very nature and purpose of these social networking sites or they would cease to exist” (Walder 2010). The judge ruled that “In light of the fact that the public portions of Plaintiff’s social networking sites contain material that is contrary to her claims and deposition testimony, there is a reasonable likelihood that the private portions of her sites may contain further evidence such as information with regard to her activities and enjoyment of life, all of which are material and relevant to the defense of this action” (Walder 2010). With social media, individuals understand portions of their personal information may be observed by others but that most people do not contemplate a comprehensive mapping of their lives over a span of weeks or months. Yet, this is exactly what happens with social media when we submit the most personal details of our lives to public scrutiny *voluntarily*.

In the Jones case, Supreme Court Justice Sotomayor suggested that the Court’s rulings that a person “Has no reasonable expectation of privacy in information voluntarily disclosed to third parties” were “Ill suited to the digital age” (Liptak 2012). She wrote “People disclose the phone numbers that they dial or text to their cellular providers; the URLs that they visit and the e-mail addresses with which they correspond to their Internet service providers; and the books, groceries, and medications they purchase to online retailers. I for one doubt that people would accept without complaint the warrantless

disclosure to the government of a list of every web site they had visited in the last week, month, or year” (Liptak 2012). Clearly a fine line of legality exists between information we voluntarily divulge to the public and parts of that information which we actively seek to protect. Regardless of the argument as to whether the information can “legally” be used, we must understand in the current digital age there is a very small “reasonable expectation” of information privacy.

Electronic mail has completely transformed the way in which we communicate with each other in the digital age and provided a vast amount of big data for organizations that store e-mail communications. This provides an enormous challenge to laws that were written to govern and protect our use of paper documents. For example, electronic records can be stored indefinitely and retrieved from electronic storage in a variety of locations. Through the use of technologies such as “keystroke loggers,” it is possible to read the contents of an e-mail regardless of whether it is ever sent. These technologies introduce a wide range of ethical issues regarding how they are used to protect or violate our personal privacy. Momentum has been building for Congress to amend the ECPA so that the law protects reasonable expectations of privacy in e-mail messages. There is a good probability that some techniques such as keystroke loggers used by employers to monitor e-mail violate the Wiretap Act. As employees increasingly communicate through instant messaging and social networks, the interception of these forms of communication also fall into question (Burgunder 2011, p. 465).

Privacy Laws

The fundamental purpose of the Fourth Amendment is to safeguard the privacy and security of individuals against arbitrary invasions by government officials. In the past, courts have affirmed that telephone calls and letters are protected media that should only be available to law enforcement with an appropriate warrant. In the same manner, electronic e-mail should be protected accordingly.

The SCA provides some level of protection for this type of data depending on the length of time the e-mail has been stored but allows subpoenas and court orders to be issued under much lower standards than those of the Fourth Amendment and provides less protection to electronic communications than wire and oral communications. While e-mail should be afforded the same level of constitutional protection as traditional forms of communication, the expectation of privacy that an individual can expect in e-mail messages depends greatly on the type of message sent and to whom the message was sent.

On December 14th, 2010, the Sixth Circuit Court of Appeals was the first and only federal appellate court to address the applicability of Fourth Amendment protection to stored emails in the landmark case of *United States v. Warshak*. The Sixth Circuit held that the “reasonable expectation” of privacy for communications via telephone and postal mail also extended to stored e-mails (Benedetti 2013, p. 77). While this was an important first step in determining the future of e-mail privacy, there still remain critical questions pertaining to the government’s ability to search and seize stored electronic communications and the proper balance between law enforcement’s need to investigate criminal activity and the individual’s need to protect personal privacy. Modern use of e-mail is as important to Fourth Amendment protections as traditional telephone conversations. Although the medium of communication will certainly change as technology evolves, the “reasonable expectation” of privacy exists in the intention of private citizens to exchange ideas between themselves in a manner that seeks to preserve the privacy of those ideas. As with the previous discussion of social media, what an individual seeks to preserve as private, even in an area accessible to the public, may be constitutionally protected (Hutchins 2007, p. 453).

Big data promises significant advances in information technology and commerce but also opens the door to a host of privacy and data protection issues. As social networking sites continue to collect massive amounts of data and

computational methods evolve in terms of power and speed, the exploitation of big data becomes an increasingly important issue in terms of cyber espionage and privacy concerns. Social media organizations have added to their data mining capabilities by acquiring other technology companies, such as when Google acquired DoubleClick and YouTube, or by moving into new fields such as Facebook did when it created “Facebook Places” (McCarthy 2010). In big data terminology, the *breadth* of an account measures the number of types of online interaction for a given user. The *depth* of an account measures the amount of data that user processes (Oboler et al. 2012). Taken together, the *breadth* and *depth* of information across multiple aspects of a user’s life can be pieced together to form a composite picture that may contain unexpected accuracy. In 2012, when Alma Whitten, Google’s Director of Privacy, Product and Engineering, announced that Google would begin to aggregate data and “treat you as a single user across all our products,” the response from users and critics was alarming. In an article for the Washington Post, Jeffrey Chester, Executive Director of the Center for Digital Democracy, voiced the reality that “There is no way a user can comprehend the implication of Google collecting across platforms for information about your health, political opinions and financial concerns” (Kang 2012). In the same article, James Steyer, Common Sense Media Chief, stated that “Google’s new privacy announcement is frustrating and a little frightening.”

Conclusion

As the world moves toward a “big data” culture centered around mobile computing, social media, and the storage of massive amounts of personal information, the threat from cyber espionage is considerable. Because of the low cost of entry and anonymity afforded by the Internet, anyone with basic technical skills can steal private information off computer networks. Due to the vast number of network attack methods and cyber espionage techniques, it is difficult to determine one effective solution. However, from the

standpoint of economics and national security, we must strive to develop a more comprehensive set of legislation and protection for the vast stores of private information readily accessible on the Internet today.

Further Reading

- Benedetti, D. (2013). How Far Can the Government’s Hand Reach Inside Your Personal Inbox? *The John Marshall Journal of Information Technology & Privacy Law – Volume 30 (Issue 1)*. Retrieved from: <http://repository.jmls.edu/cgi/viewcontent.cgi?article=1730&context=jitpl>.
- Burgunder, L. B. (2011). *Legal aspects of managing technology* (5th ed.). Mason: South-Western Cengage Learning.
- Hutchins, R. M. (2007). Tied up in knots? GPS technology and the fourth amendment. *UCLA Law Review*. Retrieved from: <http://www.uclalawreview.org/pdf/55-2-3.pdf>.
- Kang, C. (2012). Google announces privacy changes across products; users can’t opt out. *Washington Post* (24 January). Retrieved from: http://www.washingtonpost.com/business/economy/google-tracks-consumers-across-products-users-cant-opt-out/2012/01/24/gIQAArgJHOQ_story.html.
- Liptak, A. (2012). Justices say GPS tracker violated privacy rights. *New York Times*. Retrieved from: http://www.nytimes.com/2012/01/24/us/police-use-of-gps-is-ruled-unconstitutional.html?pagewanted=all&_r=0.
- McCarthy, C. (2010). Facebook granted geolocation patent. *CNet News* (6 October). Retrieved from: http://news.cnet.com/8301-13577_3-20018783-36.html.
- Oboler, A., Welsh, K., & Cruz, L. (2012). The danger of big data: Social media as computational social science. *First Monday Peer Reviewed Journal*. Retrieved from: <http://firstmonday.org/ojs/index.php/fm/article/view/3993/3269#p4>.
- Sigholm, J., & Bang, M. (2013). Towards offensive cyber counterintelligence. 2013 *European intelligence and security informatics conference*. Retrieved from: http://www.ida.liu.se/~g-johsi/docs/EISIC2013_Sigholm_Bang.pdf.
- Swingle, H. (2012). Smartphone searches incident to arrest. *Journal of the Missouri Bar*. Retrieved from: <https://www.mobar.org/uploadedFiles/Home/Publications/Journal/2012/01-02/smartphone.pdf>.
- Turley, J. (2011). Supreme court’s GPS case asks: How much privacy do we expect? *The Washington Post*. Retrieved from: http://www.washingtonpost.com/opinions/supreme-courts-gps-case-asks-how-much-privacy-do-we-expect/2011/11/10/gIQAN0RzCN_story.html.
- Walder, N. (2010). Judge grants discovery of postings on social media. *New York Law Journal*. Retrieved from: http://www.law.com/jsp/article.jsp?id=1202472483935&Judge_Grants_Discovery_of_Postings_on_Social_Media.

Cyberinfrastructure (U.S.)

Ernest L. McDuffie

The Global McDuffie Group, Longwood, FL,
USA

Introduction and Background

As the first two decades of the twenty-first century come to a close, an ever increasingly accelerated pace of technological advances continue to reshape the world. In this entry, a brief look at a number of ongoing cyberinfrastructure research projects, funded by various agencies of the United States federal government, are examined. While there are many definitions for the term cyberinfrastructure, a widely accepted one is the combination of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high-performance networks. Over the years, related research and development funding has originated more and more from government and less and less from the private sector sources. High tech companies focus mainly on applied research that can be used in new products for increased profits. This process is well suited for bringing the benefits of mature technology to the public but if allowed to become the sole destination for research funding, non-applied research will be under funded. Without basic research advances, applied research dries up and comes to an abrupt end.

A subcommittee under the Office of Science and Technology Policy (OSTP) called the National Coordinating Office (NCO) for Networking and Information Technology Research and Development (NITRD) has for almost three decades annually published a Supplement to the President's budget. (See <https://www.nitrd.gov/> for all current and past supplements.) These Supplements highlight a number of technologies. Tremendous potential impact on individuals and even greater potential impact on the nature and capabilities of the global cyberinfrastructure are clear.

Beyond these mainstream technologies is a set of fields of technological interest with even greater potential for revolutionary change. The combination of these sets, with the always occurring unpredictability of technological and scientific advances, results in significantly increased probability of world changing positive or negative events. Society can mitigate disruption caused by ignorance and the inevitable need for continual workforce realignment by focusing on the technical education of the masses. For the twenty-first century and beyond, knowledge and operational skills in the areas of science, technology, engineering, and mathematics (STEM) is critical for individuals and society.

Current Technological State

Looking at the funding picture at the federal level, impactful areas of research are for the most part computer based. All the following data comes from the 2018 NITRD Supplement which presents information based on the fiscal year 2019 federal budget request. In the Program Component Area (PCA) called Computing-Enabled Human Interaction, Communication, and Augmentation (CHuman) PCA, with significant funding request from seven federal agencies – Department of Defense (DoD), National Science Foundation (NSF), National Institutes of Health (NIH), Defense Advanced Research Projects Agency (DARPA), National Aeronautics and Space Administration (NASA), National Institute of Standards and Technology (NIST), and National Oceanic and Atmospheric Administration (NOAA) – one of its strategic priorities is human-automation interaction. This research area focuses on facilitating the interaction between humans and intelligent systems such as robots, intelligent agents, autonomous vehicles, and systems that utilize machine learning. Three of the key programs are (1) Robust Intelligence where support and the advancement of intelligent systems that operate in complex, realistic contexts is the focus; (2) Smart and Autonomous Systems research that robustly think, act, learn, and behave ethically; and (3) Smart and Connected

Communities where techno-social dimensions and their interactions in smart community environments are addressed.

In the PCA for Computing-Enabled Networked Physical Systems (CNPS), the integration of cyber/information, physical, and human worlds is accomplished using information technology-enabled systems. Twenty-one federal agencies are active in this space. They are managed by Interagency Working Groups (IWG). One is called the Cyber-Physical Systems (CPS) IWG that includes the Smart Cities and Communities Task Force, and another is named the High Confidence Software and Systems (HCSS) IWG. Research activities in these groups include investigations into cyber-physical systems, the Internet of Things (IoT), and related complex, high-confidence, networked, distributed computing systems. In the HCSS IWG, one of the key programs executing under the strategic priority of assured autonomous and artificial intelligence (AI) technologies is AI and machine learning (AI/ML) for safety and mission-critical applications. Activity here is supported by NIST, NSA, and NSF. Their efforts cover the search for techniques for assuring and engineering trusted AI-based systems, including development of shared public datasets and environments for AI/ML training, testing, and development standards and benchmarks for assessing AI technology performance.

The highest funded PCA request is High-Capability Computing Infrastructure and Applications (HCIA). Seven federal agencies – DoD, NSF, NIH, Department of Energy (DOE), DARPA, NASA, and NIST – participate in HCIA. Here the focus is on computation and data-intensive systems and applications, directly associated software, communications, storage, and data management infrastructure, and other resources supporting high-capability computing. All activities are coordinated and reported through the High End Computing (HEC) IWG which has eight participating federal agencies and strategic priorities that include, but not limited to, High-Capability Computing Systems (HCS) infrastructure as well as productivity and broadening impact.

Some of the key programs ongoing in the HEC IWG include an effort designed for the advancement of HCS applications. This research ranges

from the more basic or pure form such as applied mathematics and algorithms, and initiate activity in machine learning to optimize output from data-intensive programs at DOE; to more applied research in support of multiscale modeling of biomedical processes for improved disease treatment at NIH and multi-physics software applications to maintain military superiority by DoD. At the same time, work continues on HCS infrastructure. For example, NIH provides shared interoperable cloud computing environment, high-capacity infrastructure, and computational analysis tools for high-throughput biomedical research. A joint program demonstrating multiagency collaboration on the Remote Sensing Information Gateway is being operated by the Environmental Protection Agency (EPA), NASA, and NOAA.

AI continues to be an important area of focus for the Intelligent Robotics and Autonomous Systems (IRAS) PCA. Here funding requested across 14 agencies exploring intelligent robotic systems R&D in robotics hardware, software design and application, machines perception, cognition and adaptation, mobility and manipulation, human-robot interaction, distributed and networked robotics, and increasingly autonomous systems. Two of the four strategic priorities for IRAS are advanced robotic and autonomous systems along with intelligent physical systems where a complex set of activities are involved. These activities include the development of validate metrics, test methods, information models, protocols, and tools to advance robot system performance and safety, and develop measurement science infrastructure to specify and evaluate the capabilities of remotely operated or autonomous aerial, ground/underground, and aquatic robotic systems. Smart and autonomous systems that robustly sense, plan, act, learn, and behave ethically in the face of complex and uncertain environments are the focus.

Key programs of this PCA include the Mind, Machine, and Motor Nexus where NSF and DoD are looking at research that supports an integrated treatment of human intent, perception, and behavior in interaction with embodied and intelligent engineered systems and as mediated by motor manipulation. The Robotic Systems for Smart Manufacturing program looks to advance

measurement science to improve robotic system performance, collaboration, agility, and ease of integration into the enterprise to achieve dynamic production for assembly-centric manufacturing being executed at NIST. NIH has a Surgical Tools, Techniques, and Systems program that does R&D on next-generation tools, technologies, and systems to improve the outcomes of surgical interventions. The U.S. Navy's Office of Naval Research has a program under this PCA that is called Visual Common Sense. Here machines are developed with the capabilities to represent visual knowledge in compositional models with contextual relations and advance understanding of scenes through reasoning about geometry, functions, physics, intents, and causality.

Two closely related PCAs are the Large-Scale Data Management and Analysis (LSDMA) and the Large-Scale networking (LSN) PCS. LSDMA reports all its activities through the Big Data IWG which has 15 federal agencies. LSN forms its own IWG with some 19 federal agencies. Together these two PCAs cover several strategic priorities such as future network development, network security and resiliency, wireless networks, the effective use of large-scale data resources, and workforce development efforts to address the shortage of data science expertise necessary to move big data projects forward.

A select few of the key programs underway for LSN and LSDMA are the development of technology, standards, testbeds, and tools to improve wireless networks. Within this effort, NSF is supporting research on beyond-5G wireless technologies for scalable experimentation. 5G is the next generation of cellular networks. Currently most systems in the United States are operating a mix of 3G and 4G, with 5G set to deliver much great speed and bandwidth enabled by millimeter capable infrastructure and software applications designed to take advantage of the greater speed and higher volume of data availability. 6G and 7G networks will build of this framework and deliver capabilities difficult to even imagine over the next decade.

NSF and DARPA are leading efforts in foundational research to discover new tools and methodologies to use the massive amount of data and

information that is available to solve difficult problems. Problems related to generating alternative hypotheses from multisource data, machine reading and automated knowledge extraction, low-resource language processing, media integrity, automated software generation and maintenance, scientific discovery and engineering design in complex application domains, and modeling of global-scale phenomena. Infrastructure and tool development will focus on enabling interoperability and usability of data to allow users to access diverse data sets that interoperate both within a particular domain and across domains.

Accelerating Technological Advance

Artificial Intelligence, quantum computing, nanotechnology, and fusion power have the potential to be the biggest game-changers in terms of accelerating the pace of technological advance over the next few decades. In addition to the efforts in the United States to advance AI, other work around the world is also moving forward. AI "holds tremendous promise to benefit nearly all aspects of society, including the economy, healthcare, security, the law, transportation, even technology itself" (<https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf>). The American AI Initiative represented a whole-of-government strategy in collaboration and engagement calling for Federal agencies to prioritize R&D investments to provide education and training opportunities to prepare the American workforce and enhance access to high-quality cyberinfrastructure and data in the new era of AI.

Meanwhile, other countries, such as China, are moving forward with large-scale projects. "Tech giants, startups, and education incumbents have all jumped in. Tens of millions of students now use some form of AI to learn. It's the world's biggest experiment on AI in education, and no one can predict the outcome" (<https://www.technologyreview.com/s/614057/china-squirrel-has-started-a-grand-experiment-in-ai-education-it-could-reshape-how-the/>).

IBM has recently produced a quantum computer that can be accessed by the public. "System

One: a dazzling, delicate, and chandelier-like machine that's now the first integrated universal quantum computing system for commercial use, available for anyone to play with" (<https://singularityhub.com/2019/02/26/quantum-computing-now-and-in-the-not-too-distant-future/>, <https://www.research.ibm.com/ibm-q/>).

There is a difference between evolutionary and incremental advances in any field. Nanotechnology provides some interesting possibilities where "evolutionary nanotechnology involves more sophisticated tasks such as sensing and analysis of the environment by nano-structures, and a role for nanotechnology in signal processing, medical imaging, and energy conversion" (<http://www.trynano.org/about/future-nanotechnology>).

Fusion, the process that stars use to generate energy, is being pursued by many nations in an attempt to solve growing energy needs. "That's exactly what scientists across the globe plan to do with a mega-project called ITER. It's a nuclear fusion experiment and engineering effort to bridge the valley toward sustainable, clean, limitless energy-producing fusion power plants of the future" (<https://www.insidescience.org/video/future-fusion-energy>).

Conclusion

It may be possible to represent the interaction between technical areas and related scientific fields of interest with a bidirectional, weighted, fully connected graph where the nodes represent the different scientific fields and the edges weight represent the amount of interaction between the connected nodes. Analysis of such a graph could provide insight into where new ideas and technologies may be emerging and where best to increase funding to produce even more acceleration of the overall process.

Global cyberinfrastructure is at the center of various emerging technologies that are in the process of making major advances. The speed and impact of these potential advances are being enabled and accelerated by the growth of scale and capabilities of the cyberinfrastructure on which they depend. This interdependency will

deepen and expand, becoming indispensable for the foreseeable future.

Further Reading

- America's Energy Future: Technology and Transformation. (2009). <http://nap.edu/12091>.
 Frontiers in Massive Data Analysis. (2013). <http://nap.edu/18374>.
 Quantum Computing: Progress and Prospects. (2019). <http://nap.edu/25196>.
 Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop. (2019). <http://nap.edu/25488>.

Cybersecurity

Joanna Kulesza

Department of International Law and
 International Relations, University of Lodz, Lodz,
 Poland

Definition and Outlook

Cybersecurity is a broad term referring to measures taken by public and private entities, aimed at ensuring the safety of online communications and resources. In the context of big data it refers to the potential threats that any unauthorized disclosure of personal data or trade secrets might have on national, local, or global politics and economics. In the context of big data, cybersecurity refers to hardware, software, and individual skills deployed in order to mitigate risks originated by online transfer and storage of data, such as encryption technology, antivirus software, and employee training. Threats originating the need to introduce enhanced cybersecurity measures include but are not limited to targeted attacks by organized groups (hackers), acting independently or employed by private entities or governments. Such attacks are usually directed at crucial state or company resources. Cybersecurity threats include also malware designed to damage hardware and/or other resources by, e.g., altering their functions or allowing for a data breach. According

to the “Internet Security Threat Report 2018” by the software company Symantec other threats include cybercrime, such as phishing or spam. New threats for cybersecurity are originated by the increased popularity of social services and mobile applications, including the growing significance of GPS data and cloud architecture. They include also the “Internet of Things” with new devices granted IP addresses and providing new kinds of information, vulnerable to attack. All that data significantly fuels the big data business, offering new categories of information and new tools to process them. It significantly impacts the efficiency of customer profiling and the effectiveness of product targeting. Effectively the need for enhanced cooperation between companies offering new services and law enforcement agencies results in a heated debate on the limits of individual freedom and privacy in the global fight for cybersecurity.

Cybersecurity and Cyberwar

The break of the twenty-first century brought an increased number and impact of international hostilities affected online. Almost every international conflict has been accompanied by its online manifestation, taking on the form of malicious software deployed in order to impair rival’s critical infrastructure or state sponsored hacker groups attacking resources crucial to the opponent. The 2008 Georgia–Russia conflict resulted in attacks on Georgian authorities’ websites, originated from Russian territory. The ongoing tension between North and South Korea lead to the Oplan 5027 – a plan for US aid in case of a North Korean attack – being stolen from Seoul in 2009, while the 2011 Stuxnet virus, designed to damage Iranian uranium enrichment facilities, was allegedly designed and deployed by Israeli and US government agencies, reflecting the long-lasting Near East conflict. Next to air, water, and ground, the cyberspace has become the fourth battleground for international conflicts.

The objects of cybersecurity threats range from material resources, such as money stored in banks, offering electronic access to their services as well

as online currencies, stolen from individuals, thorough company secrets targeted by hackers hired by competition up to critical state infrastructure, including power plants, water supplies, or railroad operating systems infected with malicious code altering their operation, bringing a direct threat to the lives and security of thousands.

Cybersecurity threats may be originated by individuals or groups acting either for their own benefit, usually a financial one, or those acting upon an order or authorization of business or governments. While some of the groups conducting attacks onto critical infrastructure claim to be only unofficial supporters of national politics, like the pro-Kremlin “Nashi” group behind the 2007 Estonia attacks, forever more states, despite officially denying confirmation on individual cases, employ hackers to enhance national security standards and, more significantly, to access or distort confidential information of others. State officials admit the growing need of increased national cybersecurity by raising their military resilience in cyberspace, training hackers, and deploying elaborate spying software, designed at state demand. USCYBERCOM and Unit 61,398 of Chinese People’s Liberation Army are subject to continuous, mutual, consequentially denied accusations of espionage. Similarly, the question of German authorities permitting “Bundestrojaner” – state sponsored malicious software, used by German police for individual surveillance of Internet telephony – has been subject to heated debate over the limits of allowed compromise between privacy, state sovereignty, and cybersecurity.

Because hostile activities online often accompany offline interstate conflict, they are being referred to as acts of “cyberwar,” although the legal qualification of international hostilities enacted online as acts of war or international aggression is disputable. Forever more online conflicts attributed to states do not reflect ones ongoing offline, just to mention the long-lasting US-China tension resulting in mass surveillance by both parties of one another’s secret resources. The subsequent attacks aimed against US-based companies and government agencies, allegedly originated from Chinese territory, codenamed by

the US intelligence “Titan Rain” (2003), “electronic Pearl Harbour,” (2007) and “Operation Aurora” (2011), resulted in a breach of terabytes of trade secrets and other data. They did not, however, reflect an ongoing armed conflict between those states neither did they result in an impairment of critical state infrastructure. Following the discussion initiated in 1948 about the prohibition of force in international relations as per Article 2 (4) United Nations Charter, in 1974 the international community denied recognizing, e.g., economic sanctions as acts of war in the United Nations General Assembly Resolution 3314 (XXIX) on the Definition of Aggression, restricting the definition to direct military involvement within foreign territory. A similar discussion is ongoing with reference to cybersecurity, with one school of thought arguing that any activity causing damages or threats similar to those of a military attack ought to be considered acts of war under international law and another persistent with the narrowest possible definition of war, excluding any activity beyond a direct, military invasion of state territory from its scope, as crucial for maintaining international peace. Whether a cyberattack, threatening the life and wellbeing of many, yet affected without the deployment of military forces, tanks or machine guns, ought to be considered an act of international aggression, or, consequentially, whether lines of computer code hold similar significance to tanks crossing national borders, is unclear, and the status of an international cyberattack as an act of international aggression remains an open question.

Cybersecurity and Privacy

As individual freedom ends where joint security begins the question of the limits of permissible precautionary and counter measures against cybersecurity threats is crucial for defining it. The information on secret US mass surveillance published in 2013, describing the operation of the PRISM, UPSTREAM, and X-Keystroke programs, used for collecting and processing individual communications data, gathered by US governmental agencies following national law

yet not intact with international privacy standards incited the debate on individual privacy in the era of global insecurity. A clear line between individual rights and cybersecurity measures must be drawn. One can be found in international law documents and practice, with the guidelines and recommendations by the United Nations Human Rights Committee setting a global minimum standard for privacy. Privacy as a human right allows each individual to have their private life, including all information about them, their home, and correspondence protected by law from unauthorized interference. State authorities are obliged as per international law to introduce legal measures effectively affording such protection and act with due diligence to enable each individual under their jurisdiction the full enjoyment of their right. Any infraction of privacy may only be based on a particular provision of law, applied in individually justified circumstances. Such circumstances include the need to protect collective interests of others, that is, the right to privacy may be limited for the purpose of protecting the rights of others, including the need to guarantee state or company security. Should privacy be limited as per national law the consent of the individual whom the restriction concerns must be explicitly granted or result from particular provisions of law applied by courts with reference to individual circumstances. Moreover, states are under an international obligation to take all appropriate measures to ensure privacy protection against infraction by third parties, including private companies and individuals. This obligation results in the need to introduce comprehensive privacy laws applicable to all entities dealing with private information that accompany any security measures. Such regulations, either included in civil codes or personal data acts, are at a relatively low level of harmonization, obliging companies, in particular ones operating in numerous jurisdictions, to take it upon themselves to introduce comprehensive privacy policies, reflecting those varying national standards and meeting their clients’ demand, while at the same time ensuring company security. Effectively the issue of corporate cybersecurity needs to be discussed.

Business Security Online

According to the latest Symantec report, electronic crime will continue to enhance, resulting in the need for a tighter cooperation between private business and law enforcement. It is private actors who in the era of big data hold the power to provide unique information on their users to authorities, be it in their fight against child pornography or international terrorism. States no longer gather intelligence through their exclusive channels alone, but rather resort to laws obliging private business to convey personal information or the contents of individual correspondence to law enforcement officials. The potential threat to individual rights generated by big data technology results also in increased users' awareness of the value their information stored in the cloud and processed in bulk holds. They require for their service providers to grant them access to their personal information stored by the operator and to have the right to decide upon what happens to information so obtained. Even though international privacy standards might be relatively easy to identify, giving individuals the right to decide on what information about them may be processed and under what circumstances, national privacy laws differ thoroughly, as some states deny to recognize the right to privacy as a part of their internal legal order. Similar problem arises when freedom of speech is considered – national obscenity and state security laws differ thoroughly, even though they are based on a uniform international standard. Effectively, international companies operating in numerous jurisdictions dealing with information generated in different states need to carefully shape their policies in order to meet national legal requirements and the needs of global customers. They also need to safeguard their own interest by ensuring the security and confidentiality of rendered services. Reacting to the incoherent national laws and growing state demands, business have produced elaborate company policies, applicable worldwide. As some argue, business policies have in some areas taken over the role of national laws, with few companies in the world having more capital and effective influence on world politics

than numerous smaller states. Effectively, company policies on cooperation with state authorities and international consumer care shape global cybersecurity landscape. Projects such as the Global Network Initiative or the UN “Protect, Respect and Remedy” Framework are addressed directly to global business, transposing international human rights standards onto company obligations. While traditionally it is states who need to transpose international law onto national regulations, binding to business, in the era of big data and global electronic communications, transnational companies need to identify their own standards of cybersecurity and consumer care, applicable worldwide.

Cybersecurity and Freedom of Speech

Cybersecurity measures undertaken by states result not only in certain limitations put on individual privacy, subject to state surveillance, but also influence the scope of freedom of speech allowed by national laws. An international standard for free speech and its limits is very hard to identify. While there are numerous international conventions in place that deal with hate speech or preventing genocide and prohibit, e.g., direct and public incitement to commit genocide, it is up to states to transpose this international consensus onto national law. Following different national policies, what one state considers to be protecting national interests another might see as inciting conflicts among ethical or religious groups. Similarly, the flexible compromise on free speech, present in international human rights law, well envisaged by Article 10 of the Universal Declaration on Human Rights (UDHR) and Article 19 of International Covenant on Civil and Political Rights (ICCPR) grants everyone the right to freedom of expression, including the right to seek, receive, and impart information and ideas of all kinds, regardless of frontiers through any media, yet puts two significant limitations on this right. It may be subject to restrictions provided by law and necessary for respect of the rights or reputations of others, for the protection of national security, of public order, public health or morals. Those broad

limitative clauses allow national authorities to limit free speech exercised online for reasons of national cybersecurity. Wikileaks editors and contributors as well as PRISM whistleblower, Edward Snowden, who ousted secret information on US National Security Agency's practices, have tested the limits of allowed free speech when confronted with national security and cybersecurity agendas. The search for national cybersecurity brings forever more states to put limits on free speech exercised online, seeing secret state information or criticism of state practice as a legitimate danger to public order and state security, yet no effective international standard can be found, leaving all big data companies on their own in the search for a global free speech compromise.

Summary

Cybersecurity has become a significant issue on national agendas. It covers broad areas of state administration and public policy, stretching from military training to new interpretations of press law and limits of free speech. As cyberthreats include both direct attacks on vulnerable telecommunication infrastructure and publications considered dangerous to public order, national authorities reach for new restraints on online communications, limiting individual privacy right and free speech. Big data generates new methods for

effective online surveillance conducted by states and private business alike. Hence any discussion on cybersecurity reflects the need to effectively protect human rights exercised online. According to the United Nations Human Rights Council, new tools granted by the use of big data for cybersecurity purposes need to be used within limits set by international human rights standards.

Cross-References

- [Cyber Espionage](#)
- [Data Provenance](#)
- [Privacy](#)

Further Reading

- Brenner, J. (2013). *Glass houses: Privacy, secrecy, and cyber insecurity in a transparent world*. New York: Penguin Books.
- Clarke, R. A., & Knake, R. (2010). *Cyber war: The next threat to national security and what to do about it*. New York: Ecco.
- Deibert, R. J. (2013). *Black code: Surveillance, privacy, and the dark side of the internet*. Toronto: McClelland & Stewart.
- DeNardis, L. (2014). *The global war for internet governance*. New Haven: Yale University Press.
- Human Rights Council. Resolution on promotion, protection and enjoyment of human rights on the Internet. UN Doc. A/HRC/20/L.13.

D

Dark Web

- [Surface Web vs Deep Web vs Dark Web](#)

Darknet

- [Surface Web vs Deep Web vs Dark Web](#)

Dashboard

Christopher Pettit and Simone Z. Leao
City Futures Research Centre, Faculty of Built
Environment, University of New South Wales,
Sydney, NSW, Australia

Synonyms

[Console](#); [Control panel](#); [Indicator panel](#); [Instrument board](#)

Definition/Introduction

Dashboards have been defined by the authors of this entry as “graphic user interfaces which comprise a combination of information and geographical visualization methods for creating metrics,

benchmarks, and indicators to assist in monitoring and decision-making.”

The volume, velocity, and variety of data being produced raise challenges in how to manipulate, organize, analyze, model, and visualize such big data in the context of technology and data-driven processes for high-performance smart cities (Thakuriah et al. [2017](#)). The dashboard provides an important role both in supporting city policy and decision-making and in the democratization of digital data and citizen engagement.

Historical and Technological Evolution of Dashboards

Control Centers

The term “dashboard” inherently has its origins in the vehicle dashboard where the driver has critical information provided to them via a series of dials and indicators. The vehicle dashboard typically provides real-time information on key metrics for which the driver needs to know in making timely decisions in navigating from A to B. Such metrics include speed, oil temperature, fuel usage, distance traveled, etc. As vehicles have advanced over the decades so have their dashboard which is now typically digital and linked to the car computer which has a growing array of sensors. This is somewhat analogous to our cities which have a growing array of sensors and information, some in real time, that can be used to formulate reports against their performance. As the concept of the

dashboard has matured over recent decades, it is important to trace its lineage and highlight key events in its evolution.

Cybersyn, developed by Stafford Beer in 1970, is one of the first examples in history attempting to bring the idea of a dashboard to the context of human organizations (Medina 2011). It was a control center for monitoring the economy of Chile, based on data about the performance of various industries. The experience was not successful due to several technological limitations associated to slow and out of sync analogue data collection, and insufficient theories and methods to combine the extensive quantities of varied data.

Benefiting from advances in computer technologies, the idea of control centers has progressed in the 1980s and 1990s. Such examples include (i) the Bloomberg Terminals (1982) designed for finance professionals monitoring of business key performance indicators like cash flow, stocks, and inventory; (ii) the ComStat platform of New York City Police (1994) for aggregating and mapping crime statistics; and (iii) the Baltimore CitiStat Room (1999) for internal government accountability using metrics and benchmarks (Mattern 2015).

The Operational Center in Rio de Janeiro, Brazil, is a more recent dashboard approach in the form of a 24 hour control center developed by IBM making use of the Internet of things in a smart city context (Kitchin et al. 2015). Launched in December 2010 as a response of the government after significant death and damage due to landslides from severe rainstorms in April of the same year, it aimed at providing real-time data and analytics for emergency management. The Operational Center monitors the city in a single large control room with access to 560 cameras, multiple sensors, and a detailed weather forecasting computer model, also connected to installed sirens and an SMS system in 100 high-risk neighborhoods across the city (Goodspeed 2015).

The Operation Center in Rio is assessed as useful for dealing with emergency situations, although it does not address the root problem of landslides associated with informal urbanization. Rio's Operational Center provides a clear example of a

dashboard that is used to tackle urban management issues with real-time data yet falls short of supporting long-term strategic planning and decision-making. Therefore, some dashboard initiatives which incorporate more aggregated and pre-analyzed information which is of value to the longer-term city planning are included in the following section.

Dashboards for Planning and Decision-Making

In the context of cities, dials and indicators can be a measure of performance against a range of environmental, social, economic criteria. For city planners, policy-makers, politicians, and the community at large, it is essential that these indicators can be measured and visualized using available data for a selected city and a dashboard can provide a window into how the city is performing against such indicators. In 2014 the International Standards Organization (ISO) published ISO37120, which is a series of indicators to measure and compare the global performance of city services and quality of life. As of writing this entry, there were 38 cities from around the world that have published their performance against some citywide indicators which are available via the World Council's City Data Open Portal (<http://open.dataforcities.org/>). This data portal is essentially a city dashboard where cities can be compared against one another and indicators. Another similar Dashboard which operates at the country and regional levels of geography is the OECD's Better Life Index (<http://www.oecdbetterlifeindex.org>). The Better Life Index reports the performance of countries and regions against 11 topics including housing, income, job, employment, education, environment, civic engagement, health, life satisfaction, safety and work-life balance. Countries and regions can be compared using maps, graphs, and textual descriptions.

At a city level, Baltimore CitiStat is among the early initiatives of a dashboard approach for planning in the USA (<http://citistat.baltimorecity.gov/>), which was influential to similar developments in other cities across America. Launched in 1999

with a control center format for government internal planning, in 2003 it was complemented with an online presence through a website of city operational statistics for planners and citizens (Mattern 2015). CitiStat tracks the day-to-day operations of agencies against key performance indicators with a focus on collaboration, problem-solving, transparency and accountability, and offers some interactivity for service delivery to citizens.

The Dublin Dashboard is another example of a city level platform extending beyond real-time data (Kitchin et al. 2016). To characterize how Dublin is performing against selected indicators, and how its performance compares to other places, data are aggregated spatially and temporally, so trends and comparisons can be made. Only to characterize “what is happening right now in Dublin,” which is a small part of the dashboard, some real-time data was harvested and presented.

The Sydney 30-Minute City Dashboard (<https://cityfutures.be.unsw.edu.au/cityviz/30-min-city/>) is an example of displaying outputs of pre-developed analysis based on real-time big data. Preprocessing of real-time data was necessary due to privacy issues associated with the smart card public transport data, the large volumes of data to be stored and processed, and the aim of the platform to display the data through a specific lens (travel time against the goal of a 30min accessibility city). Total counts and statistics are presented along with informative map and graph visualizations, allowing the characterization of specific employment centers, and comparisons among them, regarding how much they deviate from the goal of a city form that promotes travel times within 30 min.

These examples were built as a multiple layer website open to any user with access to the Internet, significantly differing from control rooms. Indeed, with the ubiquity of smartphones and the growth of networked communication and social media, the engagement of citizens in the city life has changed. The democratization of data, transparency of planning processes, and participation of citizens in the assessment or planning of the city are encouraged. Dashboards have evolved to

respond to this demand and requirement for more robust citizen engagement.

Dashboards for Citizens

There is a group of dashboards that emphasizes the use of open data feeds to display real-time data, primarily focused on a single screen or web page and having individual citizens as the target audience. Examples include the London City Dashboard (Gray et al. 2016) and CityDash for Sydney (Pettit et al. 2017). For these, the design of these dashboards was guided by data availability, the frequency of update, and potential interest to citizens. As noted by Batty (2015), in this type of dashboards, each widget may have its interpretation by each user without needing a detailed analysis. The Sydney CityDash, for example, aggregates some data feeds including air quality, weather, sun protection, market information, multimodal public transit, traffic cameras, news, and selected Twitter feeds, with frequent updates every few seconds, few minutes, or an hour, depending on the parameter.

Deviating from the single screen approach, Singapore created a multilayered dashboard as their smart city platform to integrate government agencies and to engage industry, researchers, and citizens. The Singapore Smart Nation Platform (<https://www.smartnation.sg/>) was designed with the goal to improve service delivery with the use and test of innovative technologies and to promote monitoring by and feedback from the varied users, particularly intermediated by ubiquitous smartphones.

Private businesses and service providers also started to develop dashboard applications to support everyday life in cities via personal smartphones. One growing field is the monitoring of health performance based on personal movement data recorded through wearable activity tracker devices. The dashboard component of fitness tracking technology such as Fitbit and Jawbone was found to be a significant motivation factor to users by providing metrics of “progress toward a goal” (Asimakopoulos et al. 2017). Another growing area is monitoring consumption of basic utilities such as water and energy using

smartphone applications, including some that can remotely control heat or cooling in the place of residence. AGL, for example, is a large energy and gas provider in Australia which launched a smartphone application that tracks energy consumption and also energy production for those with solar systems installed, sends alerts related to user targets, and makes future forecasts (<https://www.agl.com.au/residential/why-choose-agl/agl-energy-app>). These examples benefit from smartphones not only as a tracking and control device but also as the visualization platform.

Taxonomy of Dashboards

Using seven contemporary dashboards from across the globe mentioned in this text, a proposed taxonomy has been constructed to characterize dashboards in general. The taxonomy is presented in Table 1 and includes (1) Baltimore CitiStat,

2003; (2) Rio de Janeiro Operational Center, 2010; (3) London City Dashboard, 2012 (Fig. 1a); (4) Dublin Dashboard, 2015 (Fig. 1b); (5) Smart Nation Singapore (2015); (6) Sydney CityDash, 2016 (Fig. 1c); and (7) Sydney 30-Min City Dashboard, 2016 (Fig. 1d). Figure 1 illustrates the interfaces of some of these dashboards.

Conclusions

Numerous urban and city dashboards exist nowadays, and many more are expected to be designed and developed over the coming year. This next generation of dashboards will likely include new features, and functionality will take advantage of accompanying advances in computer technologies and city governance. Kitchin and McArdle (2017) identified six key issues on how we come to know and manage cities through urban data and city dashboards: (1) How are insights and value

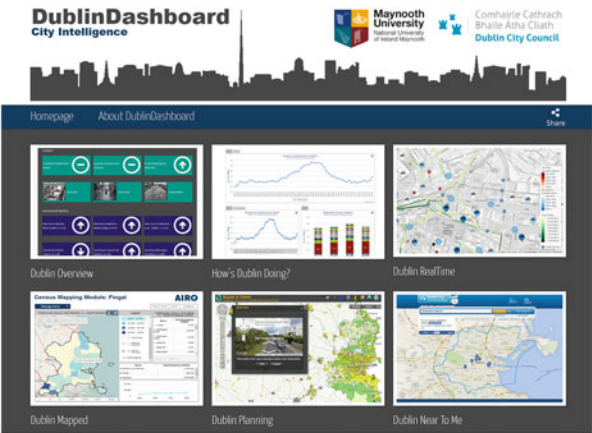
Dashboard, Table 1 Taxonomy of dashboards

Criteria	Type	Description	Examples
Access to data	Open	Uses data which is openly available to anyone, usually captured automatically by the dashboard through APIs	(1), (3), (4), (5), (6)
	Close	Uses data which is available only through licenses for specific purposes	(1), (2), (4), (5), (7)
Frequency of data	Real time	Uses data which is captured in real time by sensors and frequently updated in the dashboard through automated processing	(2), (3), (4), (5), (6)
	Preprocessed	Uses data (capture in real time or in other frequencies) which is processed and analyzed before being displayed on the dashboard	(1), (2), (4), (5), (7)
Size of data	Big data	Uses data categorized as big data, with high volume, velocity, and variety, which raises challenges for computer systems regarding data storage, sharing, and fast analysis	(1), (2), (3), (4), (5), (6), (7)
	Other	Uses data with small sizes easily managed by ordinary computer systems to store, share, and analyze data	(1), (2), (4), (5)
Dashboard audience	Decision-makers	Aims to provide information with contents, spatial and temporal scales, and analytical tools suitable to respond to urban planning issues	(1), (2), (4), (5), (7)
	Citizens	Aims to provide information with contents, spatial and temporal scales, and analytical tools suitable to respond to individual citizen issues	(1), (3), (4), (5), (6)



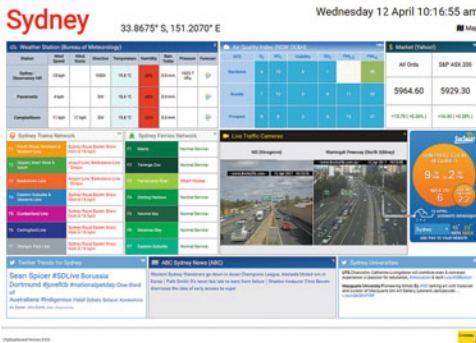
<http://citydashboard.org/london/>

(a) London Dashboard



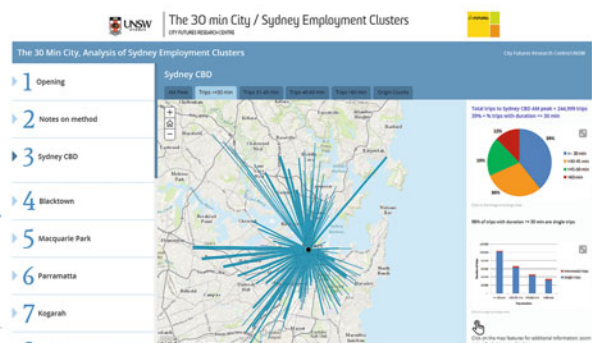
<http://www.dublindashboard.ie/pages/index>

(b) Dublin Dashboard



<http://citydashboard.be.unsw.edu.au/>

(c) Sydney CityDash



<https://cityfutures.be.unsw.edu.au/cityviz/30-min-city/>

(d) Sydney 30 Min-City Dashboard

Dashboard, Fig. 1 Examples of urban dashboards

derived from city dashboards? (2) How comprehensive and open are city dashboards? (3) To what extent can we trust city dashboards? (4) How comprehensible and useable are city dashboards? (5) What are the uses and utility of city dashboards? And (6) How can we ensure that dashboards are used ethically? Aligned to these concerns, some recommendations for the structure and design of new dashboards are suggested by Pettit et al. (2017): (1) understand the specific purpose of a dashboard and design accordingly; (2) when developing a city dashboard, human-computer interaction guidelines – particularly around usability – should be considered; (3) dashboards should support the visualization of big data to support locational insights; (4) link dashboards to established online data repositories, commonly

referred to as open data stores, clearinghouses, portals, or hubs; and (5) support a two-way exchange of information to empower citizens to engage with elements of the dashboard. A key challenge is the ability to visualize big data in real time. Incremental changes to big datasets are possible when breaking down big datasets down to individual “little” record such as tweets from a Twitter database or individual tap on and tap off records from a mass transit smart card system. However, as these systems are scaled over larger geographies, the ability to visualize big data in real time becomes more challenging. Dashboards to improve the efficiency of our cities and decision-making are an ongoing endeavor. Dashboards have proven utility in traffic management and crisis management, but when it comes to

strategic long-term planning, the value proposition of the dashboard is yet to be determined. Also, dashboards that can be used to truly empower citizens in city planning are the next frontier.

Further Reading

- Asimakopoulous, S., Asimakopoulous, G., & Spillers, F. (2017). Motivation and user engagement in fitness tracking: Heuristics for mobile healthcare wearables. *Informatics*, 2017(4), 5. <https://doi.org/10.3390/informatics4010005>.
- Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3), 274–279.
- Batty, M. (2015). A perspective on city dashboards. *Regional Studies, Region-al Science*, 2(1), 29–32.
- Goodspeed, R. (2015). Smart cities: Moving beyond urban cybernetics to tackle wicked problems. *Cambridge Journal of Regions, Economy and Society*, 8(1), 79–92.
- Gray, S., O'Brien, O., & Hügel, S. (2016). Collecting and visualizing real-time urban data through city dashboards. *Built Environment*, 42(3), 498–509.
- Kitchin, R., & McArdle, G. (2017). Urban data and city dashboards: Six key issues. In R. Kitchin, T. P. Lauriault, & G. McArdle (Eds.), *Data and the City*. London: Routledge.
- Kitchin, R., Lauriault, T. P., & McArdle, G. (2015). Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science*, 2(1), 6–28.
- Kitchin, R., Maalsen, S., & McArdle, G. (2016). The praxis and politics of building urban dashboards. *Geoforum*, 77, 93–101.
- Mattern S (2015) Mission control: A history of the urban dashboard, *Places Journal*, March 2015, <https://placesjournal.org/article/mission-control-a-history-of-the-urban-dashboard/>.
- Medina, E. (2011). *Cybernetic revolutionaries: Technology and politics in Allende's Chile*. Cambridge: The MIT Press.
- Pettit, C.J., Lieske, S., Jamal, M. (2017). CityDash: Visualising a changing city using open data. In: Geertman S, Stillwell J, Andrew, A. Pettit, C.J. (eds) Planning support systems and smart cities, lecture notes in geoinformation and cartography, *Springer International Publishing*, Basel, Switzerland, 337–353.
- Thakuriah, P. (Vonu), Tilahun, N. Y., & Zellner, M. (2017). Seeing cities through big data. *Springer Geography*. https://doi.org/10.1007/978-3-319-40902-3_1.

Data

► “Small” Data

Data Aggregation

Tao Wen

Earth and Environmental Systems Institute,
Pennsylvania State University, University Park,
PA, USA

Definition

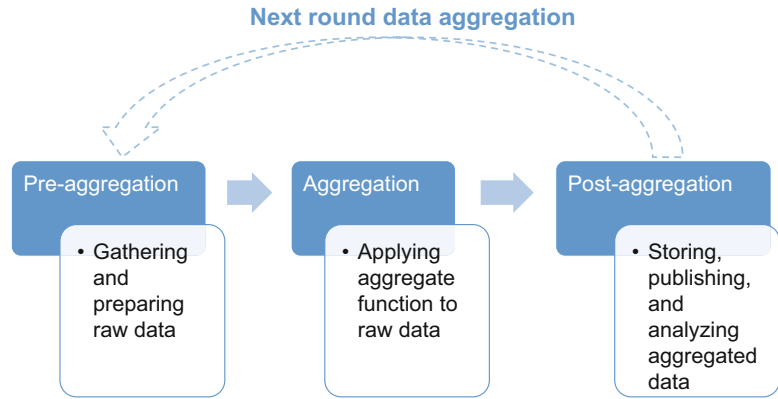
Data aggregation refers to the process by which raw data are gathered, reformatted, and presented in a summary form for subsequent data sharing and further analyses. In general, raw data can be aggregated in several ways, such as by time (e.g., monthly and quarterly), by location (e.g., city), or by data source. Aggregated data have long been used to delineate new and unusual data patterns (e.g., Wen et al. 2018). In the big data era, data are being generated at an unprecedentedly high speed and volume, which is a result of automated technologies for data acquisition. Aggregated data, rather than raw data, are often utilized to save storage space and reduce energy and bandwidth costs (Cai et al. 2019). Data aggregation is an essential component of data management, in particular during the “Analysis and Discovery” stage of the data life cycle (Ma et al. 2014).

Data Aggregation Processes and Major Issues

The processes of transforming raw data into aggregated data can be summarized as a three-step protocol (Fig. 1): (1) pre-aggregation; (2) aggregation; and (3) post-aggregation. These steps are further described below.

Pre-aggregation

This step starts with gathering data from one or more data sources. The selection of data sources is dependent on both the availability of raw data and the goal of the “Analysis and Discovery” stage. Many search tools are available to assist researchers in locating datasets and data repositories (e.g., Google Dataset Search and re3data by

Data Aggregation,**Fig. 1** General processes of data aggregation

D

DataCite). Some discipline-specific search tools are also available (e.g., DataONE for earth sciences). Data repositories generally refer to places hosting datasets. For example, Kaggle, an online repository, hosts processed datasets from a variety of disciplines. National Water Information System (NWIS) by the United States Geological Survey (USGS), and STORage and RETrieval (STORET) database by the United States Environmental Protection Agency (USEPA) both provide access to water quality data for the entire United States. Incorporated Research Institutions for Seismology (IRIS) is a collection of seismology-related data (e.g., waveform and seismic event data). Data downloaded from different sources are often not in a consistent format. In particular, data from different sources might be reported in different units (e.g., Niu et al. 2018), with different accuracy, and/or in different file formats (e.g., JavaScript Object Notation vs. Comma Separated Values). In addition, missing data are also very common. Before data aggregation at next step, data need to be cleaned and reformatted (noted as “preparing” in Fig. 1) into a unified format.

The most glaring issue in the pre-aggregation step might be related to data availability. Desired raw data might not be accessible to perform data aggregation. This situation is not uncommon, especially in business, since many of these raw data in business are considered proprietary. For example, the information of the unique identifier of persons clicking an Internet advertisement is often not accessible (Hamberg 2018). To resolve this problem, many communities, especially the

academia, start to advocate open data and FAIR Principles (i.e., findable, accessible, interoperable, and reusable) when sharing data to the data users.

Aggregation

A variety of aggregate functions are available to summarize and transform the raw data into aggregated data. These aggregate functions include (but are not limited to) minimum, mean, median, maximum, variance, standard deviation, range, sum, and count. In general, raw data can be divided into two types: numeric and categorical. Numerical data are often measurements of quantitative features (e.g., air temperature, sulfate concentration, stream discharge), and they often have mathematical meaning. Unlike numerical data, categorical data are qualitative representations (e.g., city name, mineral color, soil texture). The functions listed above might not be applicable to all types of raw data. For example, categorical data can be counted but cannot be averaged. Additionally, raw data can be aggregated over time or over space (e.g., counting the number of Fortune 500 companies in different cities). The best way to aggregate data (e.g., which aggregate function to use) should be determined by the overarching goal of the study. For example, if a researcher is interested in how housing prices fluctuate on a monthly basis for a few given cities, they should consider aggregating their raw data in two steps sequentially: (1) spatially by city, and (2) temporally aggregating data of each city by month using mean or median functions.

Data can be aggregated into groups (i.e., level of segmentation) in many different ways, e.g., housing prices of the United States can be divided by state, by county, or by city. In the aggregation step, problems can arise if raw data were not aggregated to the proper level of segmentation (Hamberg 2018). Below, an example from a water quality study is provided to illustrate this problem.

In Table 1, a hypothetical dataset of sulfate concentration (on an hourly basis) from a USGS site is listed for 3 days: 01/01/1970–01/03/1970. To calculate the mean concentration over these 3 days, a researcher should first aggregate concentration by day (each of these 3 days will have a daily mean), and then aggregate these three daily means in order to get a more representative value. Using this approach, the calculated mean sulfate concentration over these 3 days is 5 milligram/liter. Due to the fact that more sulfate measurements are available on 01/01/1970, the researcher should avoid directly aggregating these five measurements of these 3 days since this approach gives more weight on a single day, i.e., 01/01/1970. In particular, direct aggregation of these five measurements yields a biased 3-day mean of 7 milligram/liter, which is higher than 5 milligram/liter by 40%.

Post-Aggregation

In this step, aggregated data might warrant further data aggregation, in which aggregated data from last round of data aggregation will be used as the input “raw data” in the next round. Alternatively, aggregated data might be ready for data analysis, publication, and storage. For example, in the

above dataset of aggregated sulfate concentration on a monthly basis, time series analysis can be performed to determine the temporal trend of sulfate concentration, i.e., decline, increase, or unchanged.

Tools

Many tools are available for data aggregation. These tools generally fall into two categories: proprietary software and open-source software.

Proprietary software: Proprietary software is not free to use and might have less flexibility compared to open-source software; however, technical support is often more readily available for users of proprietary software. Examples of popular proprietary software include Microsoft Excel, Trifacta (Data) Wrangler, Minitab, SPSS, MATLAB, and Stata, all of which are mostly designed for preparing data (i.e., part of step 1: data cleaning and data reformatting) and aggregation (i.e., step 2). Some of these pieces of software (e.g., Excel and MATLAB) provide functions to retrieve data from varying sources (e.g., database and webpage).

Open-source software: Open-source software is free of cost to use although it might have steeper learning curve compared to proprietary software since programming or coding skills are often required to use open-source software. Open-source software can be either stand-alone program or package (or library) of functions written in free programming languages (e.g., Python and R). One example of stand-alone program is GNU Octave that is basically an open-source alternative to MATLAB, which can be used throughout all steps of data aggregation. Many programming packages are available for applications in the aggregation step (e.g., NumPy, SciPy, and Pandas in Python; dplyr and tidyr in R). These example packages can deal with data from a variety of disciplines. Some other packages including Beautiful Soup and html.parser help parse data from webpages. In certain disciplines, some packages are present to serve both steps 1 and 2, e.g., dataRetrieval in R allows users to gather and aggregate water-related data.

Data Aggregation, Table 1 Sulfate concentration (in milligram/liter; raw data) collected from 01/01/1970 to 01/03/1970 at a hypothetical USGS site

Sampling date and time	Sulfate concentration (milligram/liter)
01/01/1970, 10 AM	15
01/01/1970, 1 PM	10
01/01/1970, 4 PM	5
01/02/1970, 10 AM	2
01/03/1970, 10 AM	3

Conclusion

Data aggregation is the process where raw data are gathered, reformatted, and presented in a summary form. Data aggregation is an essential component of data management, especially nowadays when more and more data providers (e.g., Google, Facebook, National Aeronautics and Space Administration, and National Oceanic and Atmospheric Administration) are generating data at an extremely high speed. Data aggregation becomes particularly important in the era of big data because aggregated data can save storage space, and reduce energy and bandwidth costs.

Cross-References

- [Data Cleansing](#)
- [Data Sharing](#)
- [Data Synthesis](#)

Further Reading

- Cai, S., Gallina, B., Nyström, D., & Seceleanu, C. (2019). Data aggregation processes: a survey, a taxonomy, and design guidelines. *Computing*, 101(10), 1397–1429.
- Hamberg, S. (2018). Are you responsible for these common data aggregation mistakes? Retrieved 21st Aug 2019, from <https://blog.funnel.io/data-aggregation-101>.
- Ma, X., Fox, P., Rozell, E., West, P., & Zednik, S. (2014). Ontology dynamics in a data life cycle: Challenges and recommendations from a geoscience perspective. *Journal of Earth Science*, 25(2), 407–412.
- Niu, X., Wen, T., Li, Z., & Brantley, S. L. (2018). One step toward developing knowledge from numbers in regional analysis of water quality. *Environmental Science & Technology*, 52(6), 3342–3343.
- Wen, T., Niu, X., Gonzales, M., Zheng, G., Li, Z., & Brantley, S. L. (2018). Big groundwater data sets reveal possible rare contamination amid otherwise improved water quality for some analytes in a region of Marcellus shale development. *Environmental Science & Technology*, 52(12), 7149–7159.

Data Aggregators

- [Data Brokers and Data Services](#)

Data Analyst

- [Data Scientist](#)

Data Analytics

- [Business Intelligence Analytics](#)
- [Data Scientist](#)

Data Anonymization

- [Anonymization Techniques](#)

Data Architecture and Design

Erik W. Kuiler

George Mason University, Arlington, VA, USA

Introduction

The availability of Big Data sets has led many organizations to shift their emphases from supporting transaction-oriented data processing to supporting data-centric analytics and applications. The increasing rapidity of dynamic data flows, such as those generated by IoT applications and devices, the increasing sophistication of interoperability mechanisms, and the concomitant decreasing costs of data storage have transformed not only data acquisition and management paradigms but have also overloaded available ICT resources, thereby diminishing their capabilities to support organizational data and information requirements. Due to the difficulties of managing Big Data sets and increasingly more complex analytical models, transaction processing-focused ICT architectures that were sufficient to manage small data sets may require enhancements and re-purposing to support Big Data analytics.

A number of properties inform properly designed Big Data system architectures. Such architectures should be modular and scalable, able to adapt to support processing different quantities of data, and sustain real-time, high-volume, and high-performance computing, with high rates of availability. In addition, such architectures should support multitiered security and interoperability.

Conceptual Big Data System Architecture

The figure below limns a conceptual Big Data system architecture, comprising exogenous and indigenous services as well as system infrastructure services to support data analytics, information sharing, and data exchange.

Exogenous services		Indigenous services	
Access Control (Security and Privacy)	Interoperability (Data Exchange and Information Sharing)	Metadata Data Standards Data Analytics	User Delivery (Visualization and Presentation)
System Infrastructure Resource Administration; Data Storage; Orchestration; Messaging; Network Platform			

Exogenous Services

Access Control

The access control component manages the security and privacy of interactions with data providers and customers. Unlike the user delivery component, which focuses on “human” interfaces with the system, the access control component focuses on authorized access to Big Data resources via “machine-to-machine” and “human-to-machine” interfaces.

Interoperability

The interoperability component enables data exchange between different systems, regardless of data provider, recipient, or application vendor, by means of data exchange schemata and

standards. Interoperability standards, implemented at inter-system service levels, establish thresholds for exchange timeliness, transaction completeness, and content quality.

Indigenous Services

Collectively, metadata and data standards ensure syntactic conformance and semantic congruence of the contents of Big Data sets. Data analytics are executed according to clearly defined lifecycles, from context delineation to presentation of findings.

Metadata

Metadata delineate the identity and provenance of data items, their transmission timeliness and security requirements, ontological properties, etc. Operational metadata reflect the management requirements for data security and safeguarding personal identifying information (PII); data ingestion, federation, and integration; data anonymization; data distribution; and data storage. Bibliographical metadata provide information about a data item’s producer, applicable categories (keywords), etc., of the data item’s contents. Data lineage metadata provide information about a data item’s chain of custody with respect to its provenance – the chronology of data ownership, stewardship, and transformations. Syntactic metadata provide information about data structures. Semantic metadata provide information about the cultural and knowledge domain-specific contexts of a data item.

Data Standards

Data standards ensure managerial and operational consistency of data items by defining thresholds for data quality and facilitating communications among data providers and users. For example, the internationally recognized Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) provides a multilingual coded terminology that is extensively used in electronic health record (EHR) management. RxNorm, maintained by the National Institutes of Health’s National Library of Medicine (NIH

NLM), provides a common (“normalized”) nomenclature for clinical drugs with links to their equivalents in other drug vocabularies commonly used in pharmacology and drug interaction research. The Logical Observation Identifiers Names and Codes (LOINC), managed by the Regenstrief Institute, provides a standardized lexicon for reporting lab results. The International Classification of Diseases, ninth and tenth editions (ICD-9 and ICD-10), are also widely used.

Data Analytics

Notionally, data analytics lifecycles comprise a number of interdependent activities:

Context Delineation

Establishing the scope and context of the analysis define the bounds and parameters of the research initiative. Problems do not occur in vacuums; rather, they and their solutions reflect the complex interplay between organizations and the larger world in which they operate, subject to institutional, legal, and cultural constraints.

Data Acquisition

Data should come from trusted data sources and be vetted to ensure their integrity and quality prior to their preparation for analytical use.

Data Preparation

Big Data are rarely useful without extensive preparation, including integration, anonymization, and validation, prior to their use.

Data Integration Data sets frequently come from more than one provider and, thus, may reflect different cultural and semantic contexts and comply with different syntactic and symbolic conventions. Data provenance provides a basis for Big Data integration but is not sufficient by itself to ensure that the data are ready for use. It is not uncommon to give short shrift to the data integration effort and not address issues of semantic ambiguity and syntactic differences and then attempt to address these problems much later in the data analytics lifecycle, at much greater costs.

Data Anonymization The data set may contain personally identifiable information (PII) that must be addressed by implementing anonymization mechanisms that adhere to the appropriate protocols.

Data Validation Notionally, data validation comprises two complementary activities: data profiling and data cleansing. Data profiling focuses on understanding the contents of the data set and the extent to which they comply with their quality specifications in terms of accuracy, completeness, nonduplication, and representational consistency. Data cleansing focuses on normalizing the data to the extent that they are of consistent quality. Common data cleansing methods include data exclusion, to remove non-compliant data; data acceptance, if the data are within tolerance limits; data consolidation of multiple occurrences of an item; data value insertion; for example, using a default value for null fields in a data item.

Data Exploration

Data provide the building blocks with which analytical models may be constructed. Data items should be defined so that they can be used to formulate research questions and their attendant hypotheses, delineate ontological specifications and properties, identify variables (including formats and value ranges) and parameters, in terms of their interdependencies and their provenance (including chains of custody and stewardship) to determine the data’s trustworthiness, quality, timeliness, availability, and utility.

Data Staging

Because they may come from disparate sources, data may require alteration so that, for example, units of analysis are defined at the same levels of abstractions and that variables use the same code sets and are within predetermined value ranges. For example, diagnostic data from one source aggregated at the county level and the same kind of data from another source aggregated at the institutional level should be transformed so that the data conform to the same units of analysis before consolidating the data sets prior to use.

Model Development

Analytical models present interpretations of reality from particular perspectives, frequently in form of quantitative formulas that reflect particular interpretations of data sets. Induction-based algorithms may be useful, for example, in unsupervised learning settings, where the focus may be on pattern recognition, for example, in text mining, content, and topic analyses. In contrast, deduction-based algorithms may be useful in supervised learning settings, where the emphasis is on proving, or disproving, hypotheses formulated prior to analyzing the data.

Presentation of Findings

Only fully tested models should be considered ready for presentation, distribution, and deployment. Models may be presented as, for example, Business Intelligence (BI) dashboards or peer-reviewed publications. Analytics models may also be used as likelihood predictors in programmatic contexts.

User Delivery

The user delivery component presents the results of the data analytics process to end-users, supporting the transformation of data into knowledge in formats understandable to human users.

Big Data System Infrastructure

The Big Data system's infrastructure provides the foundation for Big Data interoperability and analytics by providing the components and services that support system's operations and management, from authorized, secure access to administration of system resources.

Resource Administration

The resource administration function monitors and manages the configuration, provisioning, and control of infrastructure and other components that collectively run on the ICT platform.

Data Storage

The data storage management function ensures reliable management, recording, storage of, and

access to, persistent data. This function includes logical and physical data organization, distribution, and access methods. In addition, the data storage function collaborates with metadata services to support data discovery.

Orchestration

The orchestration function configures the various architectural components and coordinates their execution so that they function as a cohesive system.

Messaging

The messaging function is responsible for ensuring reliable queuing and transmission of data and control signals between system components. Messaging may pose special problems for Big Data systems because of the computational efficiency and velocity requirements of processing Big Data sets.

Network

The network management function coordinates the transfer of data (messages) among system infrastructure components.

Platform

The ICT platform comprises the hardware configuration, operating system, software framework, and any other element on which components or services run.

Conceptual Big Data System Design Framework

User Requirements

The first activity is to determine that the research initiative meets all international, national, epistemic, and organizational ethical standards. Once this has been done, the next activity is formally to define the users' data and information requirements, the data's security and privacy requirements, and anonymization requirements. Also, users' delivery requirements (visualization and presentation) should be defined. A research question is developed to reflect the users' requirements, followed a formulation of how the

research question can be translated into a set of quantifiable, testable hypotheses.

System Infrastructure Requirements

In addition to defining user requirements, it should be determined that the system infrastructure can provide the necessary resources to undertake the research initiative.

Data Acquisition Requirements

Once the research initiative has been approved for execution, sources of the data, their security requirements, and their attendant metadata have to be identified.

Interoperability Requirements

Interoperability requirements should also be defined, for example, what data are to be shared; what standards and service level agreements (SLAs) are to be enforced; what APIs are to be used?

Project Planning and Execution

A project plan, comprising a work breakdown schedule (WBS), including time and materials allocations, should be prepared prior to project start-up. Metrics and monitoring regimens should be defined and operationalized, as should management (progress) reporting schedules and procedures.

Caveats and Future Trends

The growth and propagation of Big Data sets and their applications will continue, reflecting the impetus to develop greater ICT efficiencies and capacities to support data management and knowledge creation. To ensure the development and proper use of Big Data analytics and applications, there are, however, a number of issues that should be addressed. In the absence of an international juridical framework to govern the use of Big Data development and analytics, rules to safeguard the integrity, privacy, and security of personally identifiable information (PII) differ by country, frequently leading to confusion and the proliferation of legal ambiguities. Also, it is not

uncommon for multiple, often very different and incompatible, syntaxes, lexica, and ontologies to be in use within knowledge communities so that data may require extensive data normalization prior to their use. There are also different, competing conveyance and transportation frameworks currently in use that hamper interoperability.

Deeply troubling is the absence of a clearly defined, internationally accepted, and rigorously enforced code of ethics, with formally specified norms, roles, and responsibilities that apply to the conduct of Big Data analytics and application development. The results produced by Big Data systems have already been misused; for example, pattern-based facial recognition software is currently used prescriptively to oppress minority populations. Big Data analytics may also be misused to support prescriptive medicine without considering the risks and consequences to individuals of misdiagnoses or adverse events. In a global economy, predicated on Big Data exchanges and information sharing, developing such a code of ethics requires collaboration on epistemic, national, and international levels.

Further Reading

- NIST Big Data Public Working Group Reference Architecture Subgroup. (2015). *NIST big data interoperability framework: volume 6: Reference architecture*. Washington DC: US Department of Commerce National Institute of Standards and Technology. Downloaded from <https://bigdatawg.nist.gov>.
- Santos, M. Y., Sá, J., Costa, C., Galvão, J., Andrade, C., Martinho, B., Lima, F. V., & Costa, E. (2017). A big data analytics architecture for industry 4.0. In A. Rocha, A. Correia, H. Adeli, L. Reis, & S. Costanzo (Eds.), *Recent advances in information systems and technologies. WorldCIST 2017* (Advances in intelligent systems and computing) (Vol. 570, pp. 175–184). Cham: Springer. (Porto Santo Island, Madeira, Portugal).
- Viana, P., & Sato, L. (2015). A proposal for a reference architecture for long-term archiving, preservation, and retrieval of big data. In *13th international conference on Trust, Security and Privacy in Computing and Communications, (TrustCom)* (pp. 622–629). Beijing: IEEE Computer Society.

Data Bank

► [Data Repository](#)

Data Brokers

► [Data Mining](#)

Data Brokers and Data Services

Abdullah Alowairdhi  and Xiaogang Ma 
 Department of Computer Science,
 University of Idaho, Moscow, ID, USA

Synonyms

[Data aggregators](#); [Data consolidators](#); [Data resellers](#)

Background

Expert data brokers have been around for a long time, gathering data from media subscriptions (e.g., newspapers and magazines), mail-order retailers, polls, surveys, travel agencies, symposiums, contests, product registration and warranties, payment handling companies, government records, and more (CIPPIC 2006). In recent years, particularly since the arrival of the Internet, the data brokers' industry has expanded swiftly with the diversification of data capture and consolidation methods. As a result, a verity of products and services are offered (Kitchin 2014).

Moreover, on a daily routine, individuals involve in a variety of online activities that disclose their personal information. Such online activities include using mobile applications, buying a home or a car, subscribing to a publication, conducting a credit card transaction at department

stores or over an online catalog, participating in surveys, surfing the web, chatting with friends on a social media platform, entering sweepstakes, or subscribing to news websites. These daily activities generate a variety of information about those individuals which in turn, in many instances, is delivered or sold to data brokers (Ramirez et al. 2014).

Data Brokers

Data brokers aggregate data from a diversity of sources. In addition to existing open data sources, they also buy or rent individuals data from third-party companies. The data collected may contain activities of web browsing, bankrupt information, registrations' warranty, voting information, consumer purchase data, and other everyday web interaction activities. Typically, data brokers do not acquire data directly from individuals; hence, most individuals are unaware that their data are collected and consumed by the data brokers. Consequently, it is possible that individuals' detail life would be constructed and packaged as a final product by processing and analyzing data components supplied from different data brokers' sources (Anthes 2015).

Data brokers acquire and store individuals' data as products in a confidential data infrastructure, which stores, shares, and consumes data through networked technologies (Kitchin and Lauriault 2014). The data will be rented or sold for a profit. The data products contain lists of prospective individuals, who meet certain conditions, including details like names, telephone numbers, addresses, e-mail addresses, as well as data elements such as age, gender, income, presence of children, ethnicity, credit status, credit card ownership, home value, hobbies, purchasing habit, and background status. These derived data products collections, where data brokers have added value through data analysis and data integration methods, are used to target marketing and advertising promotions, socially classify individuals, evaluate credit rating, and tracing services (CIPPIC 2006).

Data integration and resale accompanied with correlated added value services like data analysis are a multibillion-dollar industry. Such an industry trades massive amounts of data and derived information hourly throughout a range of markets specialized in financial, retail, logistics, tourism, real estate, health, political voting, business intelligence, private security, and more. These data almost cover all aspects of everyday life and include public administration, communications, consumption of goods and media, travel, leisure, crime, and social media interactions (Kitchin 2014).

Data Sources

Selling data to brokers has become a major revenue stream for many companies. For example, retailers regularly sell data regarding customers' transactions such as credit card details, customers' purchases information and loyalty programs, customers' relationship management, and subscription information. Internet stores sell clickstream data concerning how a person navigated through a website and the time spent on different pages. Similarly, media companies, such as newspapers, radio, and television stations, gather the data contained within their content (e.g., news stories and advertisements). Likewise, social media companies aggregate the metadata and contents of their users in order to process that information to build individuals' profiles and produce their own data products to be sold to data brokers. For example, Facebook uses user networks, users' uploaded content, and user profiles of its millions of active users. The collected data, such as users' comments, videos, photos, and likes, are used to form a set of advertising products like "Lookalike Audiences, Partner Categories, and Managed Custom Audiences." Such advertising products also partner with well-known data brokers and data marketers such as Acxiom, Datalogix, Epsilon, and BlueKai in order to integrate their non-Facebook purchasing and behavior data (Venkatadri et al. 2018).

In various ways, then, individuals are handing over their own data, knowingly or unknowingly, in many volumes as subscribers, buyers, registrants, credit card holders, members, contest entrants, donors, survey participants, and web inquirers (CIPPIC 2006). Moreover, because creating, managing, and analyzing data is a specialized task, many firms subcontract their data requirements to data processing and analytics companies. By offering the same types of data services across clients, such companies can create extensive datasets that can be packaged and utilized to produce newly constructed data which provide further insights than any single source of data. In addition to these privately obtained data, data brokers also collect and consolidate public datasets such as census records, aggregate spatial data such as properties, and rent or buy data from charities and non-governmental organizations.

Methods for Data Collection and Synthesis

Data brokers aggregate data from different sources using various methods. Firstly, data brokers use crawlers and scrapers (software that extracts values from the websites and transfers these values to the data broker's data storage) to assemble publicly accessible web-based data. As an example, data brokers use software like Octoparse and import.io to decide what websites should be crawled and scraped, what data elements in each website to harvest, and how frequently. Secondly, data brokers obtain and process printed information from local government records and telephone book directories and then either process these documents using OCR (optical character recognition) scanner to produce digital records or employ specialists in data entry to create digital records manually. Thirdly, using daily data feeds, data brokers coordinate for a batch collection of data from various sources. Lastly, data brokers approach regularly to data sources through an API (application programming interface) which allow data stemming to

the data brokers' infrastructure. Whatsoever the method, data brokers may accumulate excessive data beyond their need, which may lead to the situation that they cannot attain a subset of data elements they demand. For example, some data sources sell a massive of data elements as part of a fixed dataset deal although the data broker does not request all of these data elements. Consequently, the data broker will utilize those extra data elements in some other way, such as matching or authentication purposes or to build models for new topics (Ramirez et al. 2014).

Data Markets

Data brokers construct an immense relational data infrastructure by assembling data from a variety of sources. For example, Epsilon is claimed to own loyalty card memberships data of 300 million global companies, with a database holding data related to 250 million individuals in the United States alone (Kitchin 2014). Acxiom is declared to have assembled a databank for approximately 500 million global active individuals (in the United States, around 126 million households and 190 million individuals), with around 1500 information elements per individual. Every year Acxiom handles more than 50 trillion data transactions. As a result, its income surpassing one billion dollars (Singer 2012). Moreover, it also administers separate client databases for 47 of the Fortune 100 enterprises (Singer 2012). In another example, Datalogix claims to collect data relating to offline purchases worth of trillion dollars (Kitchin 2014). Other data brokers and analysis firms including Alliance Data Systems, TransUnion, ID Analytics, Infogroup, CoreLogic, Equifax, Seisint, Innovis, ChoicePoint, Experian, Intelius, Recorded Future, and eBureau all have their unique data services and data products. For example, eBureau evaluates prospective clients on behalf of credit card companies, lenders, insurers, and educational institutions, and Intelius provides people-search services and background checks (Singer 2012).

In general, data broker companies want a wide variety of data, as large segment population as possible, that are highly relational and indexical in nature. The more data a broker can retrieve and integrate, the more likely their products work optimally and effectively, and they obtain a competitive advantage over their competitors. By gathering data together, analyzing, and organizing them appropriately, data brokers can create derived data and individual and area profiles and undertake predictive modeling to analyze individual's behavior under different situations and in different areas. This allows the more effective identification of targeted individual and provides an indication of the individual's behavior in order to reach a predetermined answer, e.g., choosing and buying specific items. Acxiom, for example, seeks to merge and process mobile data, offline data, and online data so as to generate a complete view of individual to form comprehensive profiles and solid predictive models (Singer 2012). Such information and models are very beneficial to companies because they are empowered to focus their marketing and sales efforts. The risk mitigation data products increase the possibility of successful transactions and decrease expenses relating to wastage and loss. By utilizing such products, companies thus aim to be more effective and competent in their operations.

The Hidden Business

Amusingly, slight serious attention has been paid to data brokers' operations, given the sizes and variety of individual data that they have, and how their data products are utilized to socially sort and target individuals and households. Definitely, there are a lack of academic research and media coverage regarding the consequences of data brokers' work and products. This is partly because the data broker industry is somewhat out of focus and concealed, not wanting to draw public attention, and weaken public trust in their data assets and activities, which might trigger public awareness campaigns for accountability, regulation, and

transparency. Currently, data broker industry is unregulated and is not compulsory to supply individuals with entry to their detained data. In addition, data brokers are not compelled to correct individuals' data errors (Singer 2012). Yet these data products could have a harmful reflective consequence on the services and opportunities provided to those individuals, such as whether a job will be offered, a credit application will be approved, an insurance policy will be issued, or a tenancy approved, and what price goods and services might cost based on recognized risk and value to companies (Kitchin and Lauriault 2014).

Benefits and Risks

Some benefits are feasible for individuals from data brokers' products, such as improve innovative product offerings, provide designated advertisements, and help to avoid fraud, just to name a few. The distinguished risk mitigation product delivers substantial benefits to individuals by helping avoid fraudsters from mimicking innocent individuals. Designated advertisements benefit individuals by enable them to find and enjoy the commodities and services they want and prefer more easily. Rivalry small businesses utilize data brokers' products to be able to contact certain individuals to offer innovative and improved products. However, there are numbers of possible risks from data brokers' compilation and use of individual's data. For instance, if an individual transaction is rejected because of an error in the risk mitigation product, the individual could be affected without realizing the reason. In this case, not only the individual cannot take steps to stop the problem from repeating, but also he will be deprived of the immediate benefit.

Likewise, the scoring methods used are not clear to individuals in marketing product. This means individuals are incapable of mitigating the destructive effects of lower scores. As a result, individuals are receiving inferior levels of service from companies, for example, getting limited

advertisements, or reduced subprime credit. Furthermore, marketers may use individuals' data to aid the distribution of commercial product advertisements regarding health, financial, or ethnicity, which some individuals might find disturbing and could reduce their confidence in the marketplace.

Marketers could also use the apparently harmless data inferences about individuals in ways that increase worries. For example, a data broker could be inferring that an individual classified to be in a "Speedy Drivers" data segment which will authorize a car dealership to offer that individual with a discount on sport cars. However, insurance company that uses the same data segment may deduce that the individual involves in unsafe behavior and thus will increase his insurance premium. Lastly, the people-search product can be employed to ease harassment or chase and might reveal information about victims of domestic violence, police officers, public officials, prosecutors, or other types of individuals, which might be used for revenge or other harm (Ramirez et al. 2014).

Choice as an Individual

Opt-outs are often invisible and imperfect. Data brokers may give individuals an opt-out choice for their data. Nevertheless, individuals probably do not know how to exercise this choice or even do not know the choice is presented. Additionally, individuals may find the opt-outs confusing, because the data brokers' opt-out website does not explicitly express whether the individual could utilize a choice to opt out all uses of his data. Even individuals know their data brokers websites and take the time to discover and use the opt-outs, they might still do not know its limitations. For risk mitigation products, various data brokers are not offering individuals with entry to their data or enable them to correct mistakes. For marketing products, the scope of individuals' opt-out choice is not obvious throughout their information (Ramirez et al. 2014).

Conclusion

Generally, data brokers gather data regarding individuals from a broad range of publicly available sources such as commercial and government records. The data brokers not only use the raw data collected from these sources but as well use the derived data to develop and extend their product. The three main types of products that data brokers produce for a wide range of industries are (1) people-search product, (2) marketing product, and (3) risk mitigation product. These products will be offered (i.e., sell or rent) as data packages to data brokers' clients. Several data collection methods are used by data brokers, such as web crawlers and scrapers, printed information like telephone directories, batch processing through daily feeds, and integration through an API. There are both benefits and risks for the targeted individuals in data brokers' business. Since the market of data brokers is vague, the choices to opt out the data collection are also vague. An individual needs to know the right of opt-outs to protect sensitive personal information.

Further Reading

- Anthes, G. (2015, January 1). Data brokers are watching you. Retrieved February 27, 2019, from <https://dl.acm.org/citation.cfm?doid=2688498.2686740>.
- CIPPIC. (2006). *On the data trail: How detailed information about you gets into the hands of organizations with whom you have no relationship. A report on the Canadian data brokerage industry*. Retrieved from <https://idtrail.org/files/DatabrokerReport.pdf>.
- Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. In R. Kitchin (Ed.), *Small data, data infrastructures and data brokers* (Rev. ed., pp. 27–47). London: Sage.
- Kitchin, R., & Lauriault, T. (2014, January 8). Small data, data infrastructures and big data by Rob Kitchin, Tracey Lauriault: SSRN. Retrieved February 27, 2019, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376148.
- Ramirez, E., Brill, J., Ohlhausen, M., Wright, J., & McSweeney, T. (2014). Data brokers a call for transparency and accountability. Retrieved from <https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf>.
- Singer, N. (2012, June 17). Mapping, and sharing, the consumer genome. Retrieved February 27, 2019, from <https://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html>.
- Venkatadri, G., Andreou, A., Liu, Y., Mislove, A., Gummadi, K., Loiseau, P., . . . Goga, O. (2018, May 1). Privacy risks with Facebook's PII-based targeting: Auditing a data broker's advertising interface. In *IEEE conference publication*. Retrieved February 27, 2019, from <https://ieeexplore.ieee.org/abstract/document/8418598>.

Data Center

Mél Hogan

Department of Communication, Media and Film,
University of Calgary, Calgary, AB, Canada

Synonyms

Data storage; Datacenter; Factory of the twenty-first century; Server farm; Cloud

Definition/Introduction

Big Data requires big infrastructure. A data center is largely defined by the industry as a facility with computing infrastructure, storage, and backup power. Its interior is usually designed as rows of racks containing stacked servers (a motherboard and hard drive). Most data centers are designed with symmetry in mind, alternate between warm and cool isles, and are dimly lit and noisy. The data center functions as a combination of software and hardware designed to process data requests – to receive, store, and deliver – to “serve” data, such as games, music, emails, and apps, to clients over a network. It has redundant connections to the Internet and is powered from multiple local utilities, diesel generators, battery banks, and cooling systems. Our ever-growing desire to measure and automate our world has seen a surge in data production, as Big Data.

Today, the data center is considered the heart and brain of Big Data and the Internet's networked infrastructure. However, the data center would be

defined differently across the last few decades as it underwent many conceptual and material transformations since the general-purpose computer was first imagined, and instantiated, in the 1940s.

Knowledge of the modern day data center's precursors is important because each advancement marks an important shift from elements internal to external to the apparatus, namely, in the conception of storage as memory. Computers, as we now use them, evolved from the mainframe computer as data center and today supports and serves Big Data and our digital networked communications from afar. Where and how societal data are stored has been always an important social, historical, and political question, as well as one of science and engineering, because the uses and deployments of data can vary based on context, governmental control and motivations, and level of public access.

One of the most important early examples of large-scale data storage – but which differs from today's data center in many ways – was ENIAC (Electronic Numerator, Integrator, Analyzer, and Computer), built in 1946 for the US Army Ballistic Research Laboratory, to store artillery firing codes. The installation took up 1800 sq. ft. of floor space, weighed 30 t, was expensive to run, buggy, and very energy intensive. It was kept in use for nearly a decade.

In the 1960s, there was no longer a hard distinction between processing and storage – large mainframes were also data centers. The next two decades saw the beginning and evolution of microcomputers (now called “servers”), which would render the mainframe and data center ostensibly, and if only temporarily, obsolete. Up until that point, mainframe computers used punch cards and punch tape as computer memory, which was pioneered by the textile industry for use in mechanized looms. Made possible by the advent of integrated circuits, the 1980s saw a widespread adoption of personal computers at the home and office, relying on cassette tape recorders, and later, floppy disks as machine memory. The mainframe computer was too big and too expensive to run, and so the shift to personal computing seemed to offer mitigation of these issues, which would see significant growth once again in the 1990s due to

the widespread implementation of a new client-server computing model.

Today's Data Center

Since the popularization of the public Internet in the 1990s, and especially the dot-com bubble from 1997 to 2000, data have exploded as a commodity. To put this commodity into perspective, *each minute of every day*, more than 200 million emails are sent, more than 2 million Google searches are performed, over 48 h of video is uploaded the YouTube, and more than 4 million posts appear on Facebook. Data are exploding also at the level of real-time data for services like Tinder, Uber, and AirBnB, as well as the budding self-driving car industry, smart city grids and transportation, mass surveillance and monitoring, e-commerce, insurance and healthcare transactions, and – perhaps most significantly today – the implementation of the Internet of Things (IoT), virtual and augmented reality, and gaming. All of these cloud-based services require huge amounts of data storage and energy to operate. However, despite the growing demand for storage – considering that 90% of data have been created in the last 2 years – data remain largely relegated to the realm of the ephemeral and immaterial in the public imaginary, which is a conception further upheld by the metaphor of “the cloud” and “cloud computing.” Cloud servers are no different than other data centers in terms of their materiality. They differ simply in how they provide data to users. The cloud relies on virtualization and a cluster of computers as its source to break down requests into smaller component parts (to more quickly serve up the whole) without all data (as packets) necessarily following the same physical/geographical path.

For the most part, users cannot access the servers on which their data and content are stored, which means that questions of data sovereignty, access, and ownership are also important threads in the fabric of our modern sociotechnical communication system. By foisting a guarded distance between users and their data, users are disconnected also from a proper understanding

of networked culture, and the repercussions of mass digital circulation and consumption. This distance serves companies' interests insofar as it maintains an illusion of fetching data on demand, in and from no apparent space at all, while also providing a material base that conjures up an efficient and secure system in which we can entrust our digital lives.

In reality, there are actual physical servers in data centers that contain the world's data (Neilson et al. 2016). The data center is part of a larger communications infrastructure that stores and serves data for ongoing access and retrieval. The success of the apparatus relies on uninterrupted and seamless transactions at increasingly rapid speeds. The data center can take on various forms, emplacements, and purposes; it can be imagined as a landing site (the structure that welcomes terrestrial and undersea fiber optics cables), or as a closet containing one or two locally maintained servers. But generally speaking, the data center we imagine (if we imagine one at all) is the one put on virtual display by Big Tech companies like Google, Microsoft, Facebook, Apple, Amazon, etc. (Vonderau and Holt 2015). These companies display and curate images of their data centers online and offer virtual tours to highlight their efficiency and design – and increasingly their sustainability goals and commitments to the environment. While these visual representations of data center interiors are vivid, rich, and often highly branded, the data center exteriors are for the most part boxy and nondescript. The sites are generally highly monitored, guarded, and built foremost as a kind of fortress to withstand attacks, intruders, and security breaches.

Because the scale of data centers has gotten so large, they are often referred to as server farms, churning over data, day in and day out. Buildings housing data centers can be the size of a few football fields, require millions of gallons of water daily to cool servers, and use the same amount of electricity as a midsize US town. Smaller data centers are often housed in buildings leftover and adapted from defunct industry – from underground bunkers to hotels to bakeries to printing houses to shopping malls. Data centers

(in the USA) have been built along former trade routes or railroad tracks and are often developed in the confusing context of a new but temporary market stability, itself born of economic downturns in other local industries (Burrington 2015).

Advances have been made in the last 5 years to reduce the environmental impacts of data centers, at the level energy use in particular, and this is done in part by locating data centers in locations with naturally cooler climates and stable power grids (such as in Nordic countries). The location of data centers is ultimately dependent on a confluence of societal factors, of which political stability, the risk of so-called natural disasters, and energy security remain at the top.

Conclusion

Due in part to the secretive nature of the industry and the highly skilled labor of the engineers and programmers involved, scholars interested in Big Data, new media, and networked communications have had to be creative in their interventions. This has been accomplished by drawing attention to the myth of the immaterial as a first steps to engaging every day users and politicizing the infrastructure by scrutinizing its economic, social, and environmental impacts (Starosielski 2015). The data center has become a site of inquiry for media scholars to explore and counter the widespread myths about the immateriality of “the digital” and cloud computing, its social and environmental impacts, and the political economy and ecology of communications technology more broadly.

Without denying them their technological complexities, data centers, as we now understand them, are crucial components of a physical, geographically located infrastructure that facilitates our daily online interactions on a global scale. Arguably, the initial interest in data centers by scholars was to shed light on the idea of data storage – the locality of files, on servers, in buildings, in nations – and to demonstrate the effects of the scale and speed of communication never before matched in human history. Given the rising importance of including the environment and climate change in academic and political discourse,

data centers are also being assessed for their impacts on the environment and the increasing role of Big Tech in managing natural resources. The consumption rates of water and electricity by the industry, for example, are considered a serious environmental impact because resources have, until recently, been unsustainable for the mass upscaling of its operations. Today, it is no longer unusual to see Big Tech manage forests (Facebook), partner with wastewater management plants (Google), use people as human Internet content moderators/filters (Microsoft) or own large swaths of the grid (Amazon) to power data centers. In many ways, the data industry is impacting both landscape and labor conditions in urban, suburban, rural, and northern contexts, each with its own set of values and infrastructural logics about innovation at the limits of the environment (Easterling 2014).

Cross-References

- [Big Data](#)
- [Cloud Services](#)
- [Data Repository](#)
- [Data Storage](#)
- [Data Virtualization](#)

Further Reading

- Burrington, I. (2015). How railroad history shaped internet history. *The Atlantic*, November 24. <http://www.theatlantic.com/technology/archive/2015/11/how-railroad-history-shaped-internet-history/417414>.
- Easterling, K. (2014). *Extrastatecraft: The power of infrastructure space*. London: Verso.
- Neilson, B., Rossiter, N., & Notley, T. (2016). Where's your data? It's not actually in the cloud, it's sitting in a data centre. August 30, 2016. Retrieved 20 Oct 2016, from <http://theconversation.com/wheres-your-data-its-not-actually-in-the-cloud-its-sitting-in-a-data-centre-64168>.
- Starosielski, N. (2015). *The undersea network*. Durham: Duke University Press Books.
- Vonderau, P., & Holt, J. (2015). Where the internet lives: Data centers as cloud infrastructure. In L. Parks & N. Starosielski (Eds.), *Signal traffic: Critical studies of media infrastructures*. Champaign: University of Illinois Press.

Data Cleaning

- [Data Cleansing](#)

Data Cleansing

Fang Huang
Tetherless World Constellation, Rensselaer
Polytechnic Institute, Troy, NY, USA

Synonyms

[Data cleaning](#); [Data pre-processing](#); [Data tidying](#); [Data wrangling](#)

Introduction

Data cleansing, also known as data cleaning, is the process of identifying and addressing problems in raw data to improve data quality (Fox 2018). Data quality is broadly defined as the precision and accuracy of data, which can significantly influence the information interpreted from the data (Broeck et al. 2005). Data quality issues usually involve inaccurate, unprecise, and/or incomplete data. Additionally, large amounts of data are being produced every day, and the intrinsic complexity and diversity of the data result in many quality issues. To extract useful information, data cleansing is an essential step in a data life cycle.

Data Life Cycle

“A data life cycle represents the whole procedure of data management” (Ma et al. 2014), and data cleansing is one of the early stages in the cycle. The cycle consists of six main stages (modified from Ma et al. 2014):

1. *Conceptual model*: Data science problems often require a conceptual model to define target questions, research objects, and applicable methods, which helps define the type of data to be collected. Any changes of the

conceptual models will influence the entire data life cycle. This step is essential, yet often ignored.

2. *Collection*: Data can be collected via various sources – survey (part of a group), census (whole group), observation, experimentation, simulation, modeling, scraping (automated online data collection), and data retrieval (data storage and provider). Data checking is needed to reduce simple errors and missing and duplicated values.
3. *Cleansing*: Raw data are examined, edited, and transformed into the desired form. This stage will solve some of the existing data quality issues (see below). Data cleansing is an iterative task. During stages 4–6, if any data problems are discovered, data cleansing must be performed again.
4. *Curation and sharing*: The cleaned data should be saved, curated, and updated in local and/or cloud storage for future use. The data can also be published or distributed between devices for sharing. This step dramatically reduces the likelihood of duplicated efforts. Moreover, in scientific research, open data is required by many journals and organizations for study integrity and reproducibility.
5. *Analysis and discovery*: This is the main step for using data to gain insights. By applying appropriate algorithms and models, trends and patterns can be recognized from the data and used for guiding decision-making processes.
6. *Repurposing*: The analysis results will be evaluated, and, based on the discovered information, the whole process could be performed again for the same or different target.

Data cleansing plays an essential role in the data life cycle. Data quality issues can cause extracted information to be distorted or unusable – a problem that can be mitigated or eliminated through data cleansing. Some issues can be prevented during data collection, but many have to be dealt with in the data cleansing stage. Data quality issues include errors,

missing values, duplications, inconsistent units, inaccurate data, and so on. Methods for tackling those issues will be discussed in the next sections.

Data Cleansing Process

Data cleansing deals with data quality issues after data collection is complete. The data cleansing process can be generalized into “3E” steps: examine, explore, and edit. Finding data issues through planning and examining is the most effective approach. Some simple issues like inconsistent numbers and missing values can be easily detected. However, exploratory analysis is needed for more complicated cases. Exploratory analyses, such as scatter plots, boxplots, distribution tests, and others, can help identify patterns within a dataset, thereby making errors more detectable. Once detected, the data can be edited to address the errors.

Examine

It is always helpful to define questionable features in advance, which include data type problems, missing or duplicate values, and inconsistency and conflicts. A simple reorganization and indexing of the dataset may help discover some of those data quality issues.

- *Data type problems*: In data science, the two major types of data are categorical or numeric. Categorical values are normally representation of qualitative features, such as job titles, names, nationalities, and so on. Occasionally, categorical values need to be encoded with numbers to run certain algorithms, but these remain distinct from numeric values. Numeric values are usually quantitative features, which can be further divided into discrete or continuous types. Discrete numeric values are separate and distinct, such as the population of a country or the number of daily transactions in a stock market; and continuous numeric values are usually continuous

with decimals, such the index of a stock market or the height of a person. For example, the age column of a census contains discrete numeric values, and the name column contains categorical data.

- *Missing or duplicate values*: These two issues are easily detected through reorganizing and indexing the dataset but can be hard to repair. For duplicate values, simply removing duplicated ones can solve the problem. Missing data can be filled by checking the original data records or metadata, when available. Metadata are the supporting information of data, such as the methods of measurement, environmental conditions, location, or spatial relationship of samples. However, if the required information is not in the metadata, some exploratory analysis algorithms may help fill in the missing values.
- *Inconsistency and conflicts*: Inconsistency and conflicts happen frequently when merging two datasets. Merging data representing the same samples or entities in different formats could easily cause duplication and conflicts. Occasionally, the inconsistency may not be solvable in this stage. It is acceptable to flag the problematic data and address them after the “analysis and discovery” stage, once a better overview of the data is achieved.

Explore

This stage is to use exploratory analysis methods to identify problems that are hard to find by simple examination. One of the most widely used exploratory tools is visualization, which will improve our understanding of the dataset in a more direct way. There are several common methods to do exploratory analysis on one-dimensional, two-dimensional, and multidimensional data. The dimension here refers to the number of features of data. One- and two-dimensional data can be analyzed more easily, but multidimensional data are usually reduced to lower dimensions for easier analysis and visualization. Below is a partial list of the methods that can be used in the Explore step.

One-Dimensional

- *Boxplot*: A distribution of one numeric data series with five numbers (Fox 2018). The box shows the minimum, first quartile, median, third quartile, and maximum values, allowing any outlier data points to be easily identified.
- *Histogram*: A distribution representation of one numeric data series. The numeric values are divided into bins (x-axis), and the number of points in each bin is counted (y-axis). X and Y axes are interchangeable. Its shape will change with the bin size, offering more freedom than a boxplot.

Two-Dimensional

- *Scatter plot*: A graph of the relationship between two numeric data series, whether linear or nonlinear.
- *Bar graph*: A chart to present the characteristics of categorical data. One axis represents the categories, and the other axis is the values associated with each category. There are also grouped and stacked bar graphs to show more complex information.

Multidimensional

- *Principal component analysis (PCA)*: A statistical algorithm to analyze the correlations among multivariant numeric values (Fox 2018). The multidimensional data will be reduced with two orthogonal components because it is much easier to explore the relationship among data features on a two-axis plane.

There are many other visualization and non-visualization methods in addition to the above. Data visualization techniques are among the most popular ways to identify data quality problems – they allow recognition of outliers as well as relationships within the data, including unusual patterns and trends for further analysis.

Edit

After identification of the problems, researchers need to decide how to tackle those problems, and there are multiple methods to edit the dataset. (1) Data types need to be adjusted for consistency. For instance, revise wrongly inputted numeric data to the correct values or convert numeric or categorical data to meet the requirements of the selected algorithm. (2) One should fill in the missing values and replace or delete duplicated values using the information in metadata (see *Examine* section). For example, a scientific project called “Census of Deep Life” collected microbial life samples below the seafloor along with environmental condition parameters, but some pressure values were missing. In this case, the missing pressure values were calculated using depth information recorded in metadata. (3) For inconsistency and conflicts, data conversion is needed. For example, when two datasets have different units, they should be converted before merging. (4) Some problems cannot be solved with the previous techniques and should be flagged within the dataset. In future analyses, those points can be noted and dealt with accordingly. For example, random forest, a type of machine learning algorithm, has the ability to impute missing values with existing data and relationships.

Overview

Before and during the data cleansing process, some principles should be kept in mind for best results: (1) planning and pre-defining are critical – it will give targets for the data cleansing process. (2) Use proper data structures to keep data organized and improve efficiency. (3) Prevent data problems in collection stage. (4) Use unique IDs to avoid duplication. (5) Keep a good record of metadata. (6) Always keep copies before and after cleansing. (7) Document all changes.

Tools

Many tools exist for data cleansing. There are two primary types: data cleansing software and programming packages. Software is normally easier to

use, but with less flexibility; and the programming packages have a steeper learning curve, but they are free of cost and can be extremely powerful.

- *Software*: Examples of famous software includes OpenRefine, Trifacta (Data) Wrangler, Drake, TIBCO Clarity, and many others (Deoras 2018). They often have built-in workflows and can do some statistical analysis.
- *Programming packages*: Packages written in free programming languages, such as Python and R, are becoming more and more popular in the data science industry. Python is powerful, easy to use, and runs in many different systems. The Python development community is very active and has created numerous data science libraries, including Numpy, Scipy, Scikit-learn, Pandas, Matplotlib, and so on. Pandas and Matplotlib have very powerful and easy functions to analyze and visualize different data formats. Numpy, Scipy, and Scikit-learn are used for statistical analysis and machine learning training. R is a structural programming language, similar to Python, which also has a variety of statistical packages. Some widely used R packages include dplyr, foreign, ggplot2, and tidyr, all of which are useful in data manipulation and visualization.

Conclusion

Data cleansing is essential to ensure the quality of data input for the analytics and discovery, which will in turn extract the appropriate and accurate information for future plans and decisions. This is particularly important when large tech companies, like Facebook and Twitter, 23andMe, Amazon, and Uber, and international collaboration scientific projects, are producing huge amounts of social media, genetic information, ecommerce, travel, and scientific data, respectively. Such data from various sources may have very different formats and quality, making data cleansing an essential step in many areas of science and technology.

Further Reading

- Deoras, S. (2018). 10 best data cleaning tools to get the most out of your data. Retrieved 8 Mar 2019, from <https://www.analyticsindiamag.com/10-best-data-cleaning-tools-get-data/>.
- Fox, P. (2018). Data analytics course. Retrieved 8 Mar 2019, from <https://tw.rpi.edu/web/courses/DataAnalytics/2018>.
- Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1), 81–99.
- Ma, X., Fox, P., Rozell, E., West, P., & Zednik, S. (2014). Ontology dynamics in a data life cycle: Challenges and recommendations from a Geoscience Perspective. *Journal of Earth Science*, 25(2), 407–412.
- Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2(10), e267.

Data Consolidators

► [Data Brokers and Data Services](#)

Data Discovery

Anirudh Prabhu
Tetherless World Constellation, Rensselaer
Polytechnic Institute, Troy, NY, USA

Synonyms

[Data-driven discovery](#); [Information discovery](#); [KDD](#); [KDDM](#); [Knowledge discovery](#)

Introduction/Definition

Broadly defined, data discovery is the process of finding patterns and trends in processed, analyzed, or visualized data. The reason data discovery must be defined “broadly” is because this process is popular across domains. These patterns and trends can be “discovered” from the data using different

methods depending on the context and domain of the work.

History

Recently, the term “data discovery” has been popularized as a process in Business Intelligence, with a lot of software applications and tools aiding the user in discovering trends, patterns, outliers, clusters, etc. The data discovery process itself has a longer history that dates back to the beginning of data mining. Data mining started as a trend in the 1980s and was a process of extracting information by examining databases (under human control). Other names for data mining include knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing. In 1989, Gregory Piatetsky-Shapiro introduced the notion of knowledge discovery in databases (KDD) in the first KDD workshop. The main driving factor to define the model was acknowledging the fact that knowledge is the end product of the data-driven discovery process. Another outcome of that workshop was the acknowledgement of the need to develop interactive systems that would provide visual and perceptual tools for data analysis (Kurgan and Musilek 2006). Since then, this idea has been worked on and improved upon to evolve into the “data discovery” process we know of today.

Usage in Different Contexts

Depending on context and domain of application, the process of “discovery” changes, though the end goal is identifying patterns and trends and to gain knowledge from them.

In Business Intelligence

In Business Intelligence, “data discovery” relies more on front-end analytics. The process in this domain is to have a dashboard of some kind where descriptive statistics and visualizations are

represented to the user. The user then employs this interactive dashboard to view different datasets in order to address pertinent questions. Thus, in Business Intelligence, data discovery can be defined as “a way to let people get the facts (from data) they need to do their jobs confidently in a format that’s intuitive and available” (Haan 2016).

The five principles of data analytics in business intelligence are (Haan 2016):

Fast: Data discovery is designed to answer “immediate, spur of the moment” questions. An ideal discovery solution allows access to information from many sources whenever needed. Features supporting this solution include: quick connections to many data sources, faceting and sub-setting data as required, and updating visualizations and summary statistics accordingly.

Usable: Usability and representation of the data go hand in hand. In the business intelligence domain, the data discovery process needs to remain code-free and the interface needs to be as intuitive as possible, with drag-and-drop features that make analysis steps clear as well as provide many prestructured templates, visualizations, and workflows.

Targeted: “Data discovery isn’t meant to be a monolithic practice which is the same throughout the enterprise” (Haan 2016). It needs to be customized and optimized depending on the user’s needs.

Flexible: The data discovery tool should be flexible enough to answer quick initial results from a single dataset as well as complex questions that require subsets, views, and a combination of multiple datasets. “Data discovery can and should be applied to any department or function the tool can access data for” (Haan 2016).

Collaborative: “Data discovery is not a stand-alone process to yield results.” It is when combined with analytics processes like predictive models and interactive visualizations that its usefulness is seen. “These tools should also be considered as a gateway to improving

more formal reporting, data science, and information management activities” (Haan 2016).

In Analytical Disciplines

The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data is known as knowledge discovery. Thus, data discovery and knowledge discovery are essentially interchangeably used. “The discovery process usually consists of a set of sequential steps, which often includes multiple loops and iterations a single step.” Kurgan and Musilek (2006) survey the major knowledge discovery process models, but the authors specify a slightly different terminology. According to their views, KDDM (Knowledge Discovery and Data Mining) is the process of knowledge discovery applied to any data source. Fayyad et al. (1996) describes one of the most widely cited discovery process models. The steps for this process are as follows:

- Develop an understanding of the domain and gain the relevant prior knowledge required
- Creating a target dataset (creating a data sample to perform discovery on)
- Cleaning and Preprocessing Data
- Data Reduction and Projection (Reducing the data by selecting only the most relevant variables)
- Selecting the appropriate data mining method
- Performing exploratory analysis
- Data Mining
- Interpreting mined patterns
- Documenting or using the discovered knowledge

User Involvement and Automation

Data/knowledge Discovery is considered a field where user involvement is extremely important, since the user judges whether the patterns found are useful or not. The level of user involvement and the steps where the user controls the process change depending on the field of application.

A *fully automated* discovery system is one in which the user does not need to be involved until interpretation of the returned patterns is required. McGarry (2005) mentions the characteristics necessary for an automated discovery system are to:

- Select the most relevant parameters and variables
- Guide the data mining algorithms in selecting the most salient parameters and in searching for an optimum solution
- Identify and filter the results most meaningful to the users
- Identify useful target concepts

An example of automated discovery systems that incorporate domain knowledge from the data in order to assess the novelty of patterns can be seen in the system proposed by the Ludwig process model (McGarry 2005). The system creates a prior model that can be revised every time new information is obtained. The Ludwig definition of novelty is: “a hypothesis ‘H’ is novel, with respect to a set of beliefs ‘B’, if and only if ‘H’ is not derivable from ‘B’.” This means that if a pattern contradicts a known set of beliefs, then that pattern is considered novel.

An example of a *semi-automated* discovery system is the Data Monitoring and Discovery Triggering (DMDT) system (McGarry 2005). This system limits the user’s involvement in providing feedback to guide the system as it searches. Pattern templates of “interesting” patterns will be selected and provided to the system. The DMDT system is intended to scale-up on large datasets (which in turn may be composed of multiple datasets). Over a period of time, the pattern templates defining the interesting rules will change as the data changes and will trigger new discoveries (McGarry 2005).

Discovery in the Big Data Age

As the amount of the data in the world grows exponentially, the algorithms, models, and systems proposed for specific tasks need to

be updated to accommodate massive datasets. Begoli and Horey (2012) describe some design principles that inform organizations on effective analyses and data collection processes, system organization, and data dissemination practices.

Principle 1: Support a Variety of Analysis Methods

“Most modern discovery systems employ distributed programming, data mining, machine learning, statistical analysis, and visualizations.” Distributed computing is primarily performed with Hadoop, a software product commonly coded in Java. Machine learning and statistics are generally coded in R, Python, or SAS. SQL is often employed for data mining tasks. Therefore, it is important for the discovery architecture to support a variety of analysis environments. Work is currently being done to enable this. For example, in R and Python environments, there are packages and libraries being written to run R/Python code on Hadoop environments and to also use SQL queries to mine the available data. Similarly, R and Python also have packages that are wrappers for interactive visualization libraries like D3js (written in JavaScript), which can visualize massive datasets and interactively modify views for these visualizations for the purpose of visual analysis.

Principle 2: One Size Does Not Fit All

The discovery architecture must be able to store and process the data at all the stages of the discovery process. This becomes very difficult with large datasets. Begoli and Horey (2012) proposed that instead of storing the data in one large relational database (as has been a common practice in the past), a specialized data management system is required. According to the authors, different types of analysis techniques should be able to use intermediate data structures to expedite the process.

The source data is often in an unusable format and may contain errors and missing values. Thus, the first step would be to clean the dataset and prepare it for analysis. According to Begoli

and Horey, Hadoop is an ideal tool for this step. The Hadoop framework includes MapReduce for distributed computing and scalable storage. Hive and HBase offer data management solutions for storing structured and semistructured datasets. Once the structured and semistructured datasets are stored in the format required for analysis, they can be accessed directly by the user for the machine learning/data mining tasks.

Principles 3: Make Data Accessible

This principle focuses on the representation of data and results of the data analysis. It is important to make the results (i.e., the patterns and trends) available and easy to understand. Some of the “best practices” for presenting results are:

- *Use open and popular standards:* Using popular standards and frameworks means that there is extensive support and documentation for the required analysis. For example, if custom data visualizations were created using the D3js framework, it would be easy to produce similar visualizations for different datasets by linking the front end to a constantly updating data store.
- *Using lightweight architectures:* The term lightweight architecture is used when the software application has lesser and simpler working parts than commonly known applications of the same kind. Using lightweight architectures can simplify the creation of rich applications. When combined with the open source tools mentioned earlier, they ensure that a robust application can run on a variety of platforms.
- *Interactive and flexible applications:* Users now demand rich web-enabled APIs (Application Programming Interfaces) to download, visualize, and interact with the data. So, it is important to expose part of or all the data to the users while presenting the results of the knowledge discovery process, so that they can perform additional analysis if needed.

Research and Application Challenges

This section outlines some of the obstacles and challenges faced in the discovery process. The list is by no means exhaustive, it is simply meant to give the reader an idea of the problems faced while working in this field (Fayyad et al. 1996).

Big data: Despite most of the recent research focusing on big data, this remains one of the application challenges. Using Massive datasets means that the discovery system requires large storage and powerful processing capabilities. This discovery process also needs to use efficient mining and machine learning algorithms.

High dimensionality: “Datasets with a large number of dimensions increases the size of the search space for model introduction in a ‘combinatorially explosive’ manner” (Fayyad et al. 1996). This results in the data mining algorithm finding patterns that are not useful. Dimension reduction methods combined with the use of domain knowledge can be used to effectively identify the irrelevant variables.

Overfitting: Overfitting implies that the algorithm has modeled the training dataset so perfectly that it also models the noise specific to that dataset. In this case, the model cannot be used on any other test dataset. Cross validation, regularization, and other statistical methods may be used to solve this issue.

Assessing statistical significance: “This problem occurs when the system searches for patterns over many possible models. For example, if a system tests models at the 0.001 significance level, then on average (with purely random data), N/1000 of these models will be accepted as significant. This can be fixed by adjust the test statistic as a function of the pattern search.” (Fayyad et al. 1996).

Changing data: Constantly changing data can make previously discovered patterns invalid. Sometimes, certain variables in the dataset can also be modified or deleted. This can drastically damage the discovery process. Possible solutions include incremental methods for updating patterns and using change as a trigger for a new discovery process.

Missing and noisy data: This is one of the oldest challenges in data science. Missing or noisy data can lead to biased models and thus inaccurate patterns. There are many known solutions to identify missing variables and dependencies.

Complex relationships between attributes: In some cases, the attributes in a dataset may have a complex relationship with each other (for example, a hierarchical structure). Older machine learning/data mining algorithm might not take these relationships into account. It is important to use algorithms that derive relations between the variables and create pattern based on these relations.

Understandability of patterns: The results of the discovery process need to be easy to understand and interpret. Well-made interactive visualizations, combined with summarizations in natural language, is a good starting step to address this problem.

Integration: Discovery systems typically need to integrate with multiple data stores and visualization tools. These integrations are not possible if the tools integrated are not interoperable. Use of open source tools and framework helps address this problem.

Cross-References

► Data Processing

Further Reading

- Begoli, E., & Horey, J. (2012). Design principles for effective knowledge discovery from big data. In *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 joint working IEEE/IFIP conference on IEEE* (pp. 215–218). IEEE.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37.
- Haan, K. (2016). So what is data discovery anyway? 5 key facts for BI. Retrieved Sept 24, 2017, from <https://www.ironsidegroup.com/2016/03/21/data-discovery-5-facts-bi/>.
- Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(1), 1–24.
- McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(1), 39–61.

Data Exhaust

Daniel E. O’Leary¹ and Veda C. Storey²

¹Marshall School of Business, University of Southern California, Los Angeles, CA, USA

²J Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA

Overview

Data exhaust is a type of big data that is often generated unintentionally by users from normal Internet interaction. It is generated in large quantities and appears in many forms, such as the results from web searches, cookies, and temporary files. Initially, data exhaust has limited, or no, direct value to the original data collector. However, when combined with other data for analysis, data exhaust can sometimes yield valuable insights.

Description

Data exhaust is passively collected and consists of random online searches or location data that is generated, for example, from using smart phones with location dependent services or applications (Gupta and George 2016). It is considered to be “noncore” data that may be generated when individuals use technologies that passively emit information in daily life (e.g., making an online purchase, accessing healthcare information, or interacting in a social network). Data exhaust can also come from information-seeking behavior that is used to make inferences about an individual’s needs, desires, or intentions, such as Internet searches or telephone hotlines (George et al. 2014).

Additional Terminology

Data exhaust is also known as ambient data, remnant data, left over data, or even digital exhaust (Mcfedries 2013). A digital footprint or a digital dossier is the data generated from online activities

that can be traced back to an individual. The passive traces of data from such activities are considered to be data exhaust. The big data that interests many companies is called “found data.” Typically data is extracted from random Internet searches and location data is generated from smart or mobile phone usage. Data exhaust should not be confused with community data that is generated by users in online social communities, such as Facebook and Twitter.

In the age of big data, one can, thus, view data as a messy collage of data points, which includes found data, as well as the data exhaust extracted from web searches, credit card payments, and mobile devices. These data points are collected for disparate purposes (Harford 2014).

Generation of Data Exhaust

Data exhaust is normally generated autonomously from transactional, locational, positional, text, voice, and other data signatures. It typically is gathered in real time. Data exhaust might not be purposefully collected, or is collected for other purposes and then used to derive insights.

Example of Data Exhaust

An example of data exhaust is backend data. Davidson (2016) provides an example from a real-time information transit application called Transit App (Davidson 2016). The Transit App provides a travel service to users. The App shows the coming departures of nearby transit services. It also has information on bike share, car share, and other ride services, which appear when the user simply opens the app. The app is intended to be useful for individuals who know exactly where they are going and how to get there, but want real-time information on schedules. The server, however, retains data on the origin, destination, and device data for every search result. The usefulness of this backend data was assessed by comparing the results obtained from using the backend data to predict trips, to a survey data of actual trips, which revealed a very similar origin-destination pattern.

Sources of Data Exhaust

The origin of data exhaust may be passive, digital, or transactional. Specifically, data exhaust can be passively collected as transactional data from people’s use of digital services such as mobile phones, purchases, web searches, etc. These digital services are then used to create networked sensors of human behavior.

Potential Value

Data exhaust is accessed either directly in an unstructured format or indirectly as backend data. The value of data exhaust often is in its use to improve online experiences and to make predictions about consumer behavior. However, the value of the data exhaust can depend on the particular application and context.

Challenges

There are practical and research challenges to deriving value from data exhaust (technical, privacy and security, and managerial). A major technical challenge is the acquisition of data exhaust. Because it is often generated without the user’s knowledge, this can lead to issues of privacy and security. Data exhaust is often unstructured data for which there is, technically, no known, proven, way to consistently extract its potential value from a managerial perspective. Furthermore, data mining and other tools that deal with unstructured data are still at a relatively early stage of development.

From a research perspective, traditionally, research studies of humans have focused on data collected explicitly for a specific purpose. Computational social science increasingly uses data that is collected for other purposes. This can result in the following (Altman 2014):

1. Access to “data exhaust” cannot easily be controlled by a researcher. Although a researcher may limit access to their own data, data exhaust may be available from commercial sources or from other data exhaust sources. This increases the risk that any sensitive information linked with a source of data exhaust can be reassociated with an individual.

2. Data exhaust often produces fine-grained observations of individuals over time. Because of regularities in human behavior, patterns in data exhaust can be used to “fingerprint” an individual, thereby enabling potential reidentification, even in the absence of explicit identifiers or quasi-identifiers.

Evolution

As ubiquitous computing continues to evolve, there will be a continuous generation of data exhaust from sensors, social media, and other sources (Nadella and Woodie 2014). Therefore, the amount of unstructured data will continue to grow and, no doubt, attempts to extract value from data exhaust will grow as well.

Conclusion

As the demand for capture and use of real-time data continues to grow and evolve, data exhaust may play an increasing role in providing value to organizations. Much communication, leisure, and commerce occur on the Internet, which is now accessible from smartphones, cars, and a multitude of devices (Harford 2014). As a result, activities of individuals can be captured, recorded, and represented in a variety of ways, most likely leading to an increase in efforts to capture and use data exhaust.

Further Reading

- Altman, M. (2014). *Navigating the changing landscape of information privacy*. <http://informatics.mit.edu/blog/2014/10/examples-big-data-and-privacy-problems>.
- Bhushan, A. (2013). “Big data” is a big deal for development. In Higgins, K. (Ed), *International development in a changing world*, 34. The North-South Institute, Ottawa, Canada.
- Davidson, A. (2016). Big data exhaust for origin-destination surveys: Using mobile trip-planning data for simple surveying. *Proceedings of the 95th Annual Meeting of the Transportation Research Board*.
- George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, 57(2), 321–326.

- Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information Management*, 53(8), 1049–1064.
- Harford, T. (2014). Big data: A big mistake? *Significance*, 11(5), 14–19.
- Mcfedries, P. (2013). Tracking the quantified self [Technically speaking]. *IEEE Spectrum*, 50(8), 24–24.
- Nadella, A., & Woodie, A. (2014). Data ‘exhaust’ leads to ambient intelligence, Microsoft CEO says. https://www.datanami.com/2014/04/15/data_exhaust_leads_to_ambient_intelligence_microsoft_ceo_says/.

Data Fusion

Carolynne Hultquist
Geoinformatics and Earth Observation
Laboratory, Department of Geography and
Institute for CyberScience, The Pennsylvania
State University, University Park, PA, USA

Definition/Introduction

Data fusion is a process that joins together different sources of data. The main concept of using a data fusion methodology is to synthesize data from multiple sources in order to create collective information that is more meaningful than if only using one form or type of data. Data from many sources can corroborate information, and, in the era of big data, there is an increasing need to ensure data quality and accuracy. Data fusion involves managing this uncertainty and conflicting data at a large scale. The goal of data fusion is to create useful representations of reality that are more complete and reliable than a single source of data.

Integration of Data

Data fusion is a process that integrates data from many sources in order to generate more meaningful information. Data fusion is very domain-dependent, and therefore, tasks and the development of methodologies are dependent on the field for diverse purposes (Bleiholder

and Naumann 2008). In general, the intention is to fuse data from many sources in order to increase value. Data from different sources can support each other which decreases uncertainty in the assessment or conflicts which raises questions of validity. Castanedo (2013) groups the data fusion field into three major methodological categories of data association, state estimation, and decision fusion. Analyzing the relationships between multiple data sources can help to provide an understanding of the quality of the data as well as identify potential inconsistencies.

Modern technologies have made data easier to collect and more accessible. The development of sensor technologies and the interconnectedness of the Internet of things (IoT) have linked together an ever-increasing number of sensors and devices which can be used to monitor phenomena. Data is accessible in large quantities, and multiple sources of data are sometimes available for an area of interest. Fusing data from a variety of forms of sensing technologies can open new doors for research and address issues of data quality and uncertainty.

Multisensor data fusion can be done for data collected for the same type of phenomena. For example, environmental monitoring data such as air quality, water quality, and radiation measurements can be compared to other sources and models to test the validity of the measurements that were collected. Geospatial data is fused with data collected in different forms and is sometimes also known in this domain as data integration. Geographical information from such sources as satellite remote sensing, UAVs (unmanned aerial vehicles), geolocated social media, and citizen science data can be fused to give a picture that any one source cannot provide. Assessment of hazards is an application area in which data fusion is used to corroborate the validity of data from many sources. The data fusion process is often able to fill some of the information gaps that exist and could assist decision-makers by providing an assessment of real-world events.

Conclusion

The process of data fusion directly seeks to address challenges of big data. The methodologies are directed at considering the veracity of large volumes and many varieties of data. The goal of data fusion is to create useful representations of reality that are more complete and reliable than trusting data that is only from a single source.

Cross-References

- [Big Data Quality](#)
- [Big Variety Data](#)
- [Data Integration](#)
- [Disaster Planning](#)
- [Internet of Things \(IoT\)](#)
- [Sensor Technologies](#)

Further Reading

- Bleiholder, J., & Naumann, F. (2008). Data fusion. *ACM Computing Surveys*, 41, 1:1–1:41.
- Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, 2013, 1–19, Article ID 704504.

Data Governance

Erik W. Kuiler

George Mason University, Arlington, VA, USA

Introduction

Big Data governance is the exercise of decision-making for, and authority over, Big Data-related matters. Big Data governance comprises a set of decision rights and accountabilities for Big Data and information-related processes, executed according to agreed-to processes, standards, and models that collectively describe who can take what actions with what information and when, in accordance with predetermined methods and authorized access rights.

Distinctions Between Big Data Governance, Big Data Management, Big Data Operations, and Data Analytics

Big Data governance is a shared responsibility and depends on stakeholder collaboration so that shared decision-making becomes the norm, rather than the exception, of responsible Big Data governance.

As a component of an overall ICT governance framework, Big Data governance focuses on the decisions that must be made to ensure effective management and use of Big Data and decision accountability.

Big Data management focuses on the execution of Big Data governance decisions. The Big Data management function administers, coordinates, preserves, and protects Big Data resources. In addition, this organization is responsible for developing Big Data management procedures, guidelines, and templates in accordance with the direction provided by the Big Data governance board. The Big Data management function executes Big Data management processes and procedures, monitors their compliance with Big Data governance policies and decisions, and measures the effectiveness of Big Data operations. In addition, the Big Data management function is responsible for managing the technical Big Data architecture.

Big Data operations function focuses on the execution of the activities stipulated by Big Data management and on capturing metrics of its activities. In this context, the focus of Big Data operations is on the effective execution of the Big Data management life cycle, broadly defined as Big Data acquisition and ingestion; Big Data integration, federation, and consolidation; Big Data normalization; Big Data storage; Big Data distribution; and Big Data archival. The Big Data operations function directly supports the various Big Data user groups, including informaticists, who focus on developing Big Data analysis models (descriptive, predictive, and, where feasible, prescriptive analytics), business intelligence models, and Big Data visualization products.

Big Data Governance Conceptual Framework

The figure below depicts a conceptual framework of Big Data governance, comprising a set of integrated components.



Data Governance, Fig. 1 Data integration – different sources

Big Data Governance Foundations

Big Data governance is an ongoing process that, when properly implemented, ensures the alignment of decision makers, stakeholders, and users with the objectives of the authorized, consistent, and transparent use of data assets. In a Big Data-dependent environment, change is inevitable, and achieving collaboration among various stakeholders, such as strategic and operational managers, operational staff, customers, researchers, and analysts, cannot be managed as one-time events.

Guiding Principles, Policies, and Processes

Effective Big Data governance depends on the formulation of guiding principles and their transformation into implementable policies that are operationalized as monitored processes. Examples are: maintaining the ontological integrity of persons; a code of ethics to guide Big Data operations, applications, and analytics; a single version of the “truth” (the “golden record”);

recognition of clearly identified Big Data sources and data recipients (lineage, provenance, and information exchange); transparency; unambiguous transformation of Big Data; Big Data quality, integrity, and security; alignment of Big Data management approaches; integrity and repeatability of Big Data processes and analytics methods; Big Data analytics design reviews; systemic and continuous Big Data governance and management processes, focusing on, for example, configuration and change management, access management, data life cycle management. Processes should be repeatable and include decision points (stage gates), escalation rules, and remediation processes.

Big Data Analytics, Ethics, and Legal Considerations

The ability to process and analyze Big Data sets has caused an epistemic change in the approach to data analytics. Rather than treating data as if they are a bounded resource, current ICT-supported algorithm design and development capabilities enable data to now operate as nodes in ever-expanding global networks of ontological perspectives, each of which comprises its own set of shareable relationships.

The legal frameworks are not yet in place fully to address complexities that accompany the availability of Big Data. Currently, self-regulation sustains the regulatory paradigm of Big Data governance, frequently predicated on little more than industry standards, augmented by technical guidelines, and national security standards. Governance of such a complex environment requires the formulation of policies at the national and international level that address ethical use of Big Data applications that go beyond technical issues of identity and authentication, access control, communications protocols and network security, and fault tolerance. The ethical use of Big Data depends on establishing and sustaining complex of trust; for example, from a technological perspective, trust in the completeness of a transaction; from a human perspective, trust that an agent in the Internet of things (IoT) will not compromise

a person's ontological integrity; from a polity perspective, that Big Data applications such as Artificial Intelligence (AI), will not be used prescriptively to mark individuals as undesirable or unwelcome, based, for instance, on nothing more than cultural prejudices, gender biases, or political ideologies.

Big Data Privacy and Security

Data privacy mechanisms have traditionally focused on safeguarding Personally Identifiable Information (PII) against unauthorized access. The availability of Big Data, especially IoT-produced data, complicates matters. The advent of IoT provides new opportunities for Big Data-sustained processes and analytics. For example, medical devices in IoT environments may act as virtual agents and operators, each with its own ontological aspects of identity (beyond radio frequency identification tags) and evolutionary properties, effectively blurring the distinctions between the digital and the physical spheres of this domain and raising not only ethical questions but also questions of the sufficiency and efficacy of the governance framework to address the privacy and security requirements of managing the lifecycles of citizens' data from their capture to archival.

Lexica, Ontologies, and Business Rules

A lexicon provides a controlled vocabulary that contains the terms and their definitions that collectively constitute a knowledge domain. A lexicon enforces rule-based specificity of meaning, enabling semantic consistency and reducing ambiguity and supports the correlation of synonyms and the semantic (cultural) contexts in which they occur. Furthermore, to support data interoperability, a lexicon provides mechanisms that enable cross-lexicon mapping and reconciliation. The terms and their definitions that constitute the lexicon provide the basis for developing ontologies, which delineate the interdependencies among categories and their properties, usually in the form of similes, meronymies, and metonymies. Ontologies encapsulate the intellectual

histories of epistemic communities and reflect social constructions of reality, defined in the context of specific cultural norms and symbols, human interactions, and processes that collectively facilitate the transformation of data into knowledge. Ontologies support the development of dynamic heuristic instruments that sustain Big Data analytics. Business rules operationalize ontologies by establishing domain boundaries and specifying the requirements that Big Data must meet to be ontologically useful rather than to be excluded as “noise.”

Metadata

Metadata are generally considered to be information about data and are usually formulated and managed to comply with predetermined standards. Operational metadata reflect the requirements for data security; data anonymizing, including personally identifying information (PII); data ingestion, federation, and integration; data distribution; and analytical data storage. Structural (syntactic) metadata provide information about data structures. Bibliographical metadata provide information about data set producers, such as the author, title, table of contents, applicable keywords of a document; data lineage metadata provide information about the chain of custody of a data item with respect to its provenance – the chronology of data ownership, stewardship, and transformations. Metadata also provide information on the data storage locations, usually as either as local, external, or as cloud-based data stores.

Big Data Quality

Both little and Big Data of acceptable quality are critical to the effective operations of an organization and to the reliability of its business intelligence and analytics. Data quality is a socio-cultural construct, defined by an organization in the context of its mission and purpose. Big Data quality management is built on the fundamental premise that *data* quality is meaningful only to the extent that it relates to the intended use of the data. Juran (1999, 34.9) notes, “Data are of high quality if they are fit

for their intended uses in operations, decision making and planning. Data quality means that data are relevant to their intended uses and are of sufficient detail and quantity, with a high degree of accuracy and completeness, consistent with other sources, and presented in appropriate ways.”

Big Data Interoperability

Big Data interoperability relies on the secure and reliable transmission of data that conform to predetermined standards and conventions that are encapsulated in the operations of lexicon and ontology, metadata, and access and security components of the data governance framework. The application of Big Data interoperability capabilities may introduce unforeseen biases or instances of polysemy that may compromise, albeit unintentionally, the integrity of the research and the validity of its results. The growth of IoT-produced Big Data may exacerbate transparency issues and ethical concerns and IoT may also raise additional legal issues. For example, an IoT agent may act preemptively and prescriptively on an individual’s behalf without his or her knowledge. In effect, individuals may have abrogated, unknowingly, their control over their decisions and data. Moreover, IoT agents can share information so that data lineage and provenance chains become confused or lost, culminating in the compromise of data quality and reliability.

Big Data Analytics

Drawing liberally from statistics, microeconomics, operations research, and computer science, Big Data analytics constitute an integrative discipline to extract and extrapolate information from very large data sets. Increasingly, organizations use data analytics to support data-driven program execution, monitoring, planning, and decision-making. Data analytics provide the means to meet diverse information requirements, regardless of how they may be used to present or manage information. Big Data analytics lifecycle stages comprise a core set of data analytics capabilities, methods, and techniques that can be

adapted to comply with any organization's data governance standards, conventions, and procedures.

Challenges and Future Trends

Melvin Kranzberg (1986) observes, "Technology is neither good nor bad; nor is it neutral," as a reminder that we must constantly compare short-term with long-term results, "the utopian hopes versus the spotted reality, what might have been against what actually happened, and the trade-offs among various 'goods' and possible 'bads'" (pp. 547–548). The proliferation of Big Data and the proliferation of IoT environments will continue, requiring the development and implementation of flexible Big Data governance regimes. However, the velocity of change engendered by Big Data and IoT expansion has exacerbated the difficulties of defining and implementing effective Big Data governance programs without compromising those standards that, *à priori*, define ethical use of Big Data in cloud-based, global IoT environments.

Further Reading

- Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52(8), 36–44.
- Juran, J., & Godfrey, B. (1999). *Juran's Quality Handbook*. (Fifth Ed.). New York: McGraw-Hill.
- Kranzberg, M. (1986). Technology and history: "Kranzberg's Laws". *Technology and Culture*, 27(3), 544–560.
- National Research Council. (2013). *Frontiers in massive data analysis*. Washington, DC: The National Academies Press.
- Roman, R., Zhou, J., & Lopez, J. (2013). On the features and challenges of security and privacy in distributed network of things. *Computer Networks*, 57(10), 2266–2279.
- Sinaeepourfard A., J. Garcia, X. Masip-Bruin., & Marín-Torder E. (2016). Towards a comprehensive data lifecycle model for big data environments. In *2016 IEEE/ACM 3rd international conference on big data computing, applications and technologies* (pp. 100–106).
- Weber, R. H. (2010). Internet of things – new security and privacy challenges. *Computer Law and Security Review*, 26, 23–30.

Data Hacker

► [Data Scientist](#)

Data Integration

Anirudh Kadadi¹ and Rajeev Agrawal²

¹Department of Computer Systems Technology,
North Carolina A&T State University,
Greensboro, NC, USA

²Information Technology Laboratory, US Army
Engineer Research and Development Center,
Vicksburg, MS, USA

Synonyms

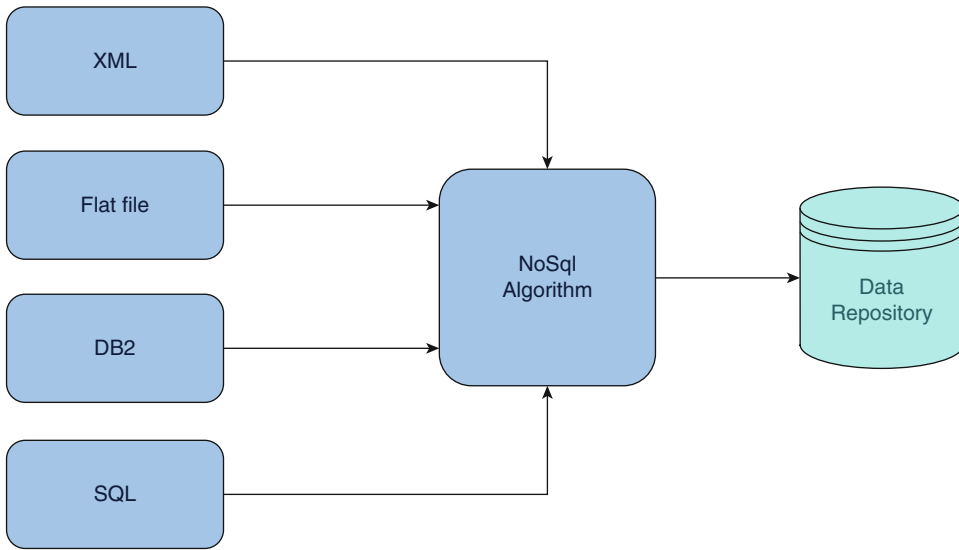
[Big data](#); [Big data integration tools](#); [Semi-structured data](#); [Structured data](#); [Unstructured data](#)

Introduction

Big data integration can be classified as a crucial part of integrating enormous datasets in multiple values. The big data integration is a combination of data management and business intelligence operations which covers multiple sources of data within the business and other sources. This data can be integrated into a single subsystem and utilized by organizations for business growth. Big data integration also involves the development and governance of data from different sources which could impact organization's abilities to handle this data in real time.

The data integration in big data projects can be critical as it involves:

1. Discovering the sources of data, analyzing the sources to gain bigger insights of data, and profiling the data.
2. Understanding the value of data and analyzing the organizational gains through this data. This can be achieved by improving the quality of data.



Data Integration, Fig. 1 Data integration – different sources

3. Finally transforming the data as per the big data environment (Fig. 1).

The five Vs of big data can influence the data integration in many ways. The five Vs can be classified as volume, velocity, variety, veracity, and value: The enormous volume of data is generated every second in huge organizations like Facebook and Google. In the earlier times, the same amount of data was generated every minute. This variation in data-generating capacity of the organizations has been increasing rapidly, and this could motivate the organizations to find alternatives for integrating the data generated in larger volumes for every second. The speed at which the data is transmitted from the source to destination can be termed as velocity. Data generated by different jobs at each time is transmitted at timely basis and stored for further processing. In this case, the data integration can be performed only after a successful data transmission to the database. The data comes from numerous sources which categorizes them into structured and unstructured. The data from social media can be the best example for unstructured data which includes logs, texts, html tags, videos, photographs, etc. The data integration in this scenario can be performed only on the relational data

which is already structured and the unstructured data has to be optimized to structured data before the data integration is performed (Vassiliadis et al. 2002).

The trustworthiness and accuracy of the data from the sources can be termed as the veracity. The data from different sources comes in the form of tags and codes where organizations were lagging the technologies to understand and interpret this data. But technology today provides us the flexibility to work with these forms of data and use it for business decisions. The data integration jobs can be created on this data depending on the flexibility and trust of this data and its source. The value can be termed as the business advantage and profits the data can bring to the organization. The value depends solely on the data and its source. Organizations target their profits using this data, and this data remains at a higher stake for different business decisions across the organization. Data integration jobs can be easily implemented on this data, but most of the organizations tend to keep this data as a backup for their future business decisions. Overall, the five V's of big data play a major role in determining the efficiency of organizations to perform the data integration jobs at each level (Lenzerini 2002).

Traditional ETL Methods with Hadoop as a Solution

Organizations tend to implement the big data methodologies into their work system creating information management barriers which include access, transform, extract, and load the information using traditional methodologies for big data. Big data creates potential opportunities for organizations. To gain the advantage over the opportunities, organizations tend to develop an effective way of processing and transforming the information which involves data integration at each level of data management. Traditionally, data integration involves integration of flat files, in-memory computing, relational databases, and moving data from relational to non-relational environments.

Hadoop is the new big data framework which enables the processing of huge datasets from different sources. Some of the market leaders are working on integrating Hadoop with the legacy systems to process their data for business use in current market trend. One of the oldest contributors to the IT industry “the mainframe” has been into existence since a long time, and currently IBM is working on development of new techniques to integrate the large datasets through Hadoop and mainframe.

The Challenges of Data Integration

In a big data environment, the data integration can lead to many challenges in real-time implementation which has the direct impact on projects. Organizations tend to implement new ways to integrate this data to derive meaningful insights at a bigger picture. Some of the challenges posed in data integration are discussed as:

(i) *Accommodate scope of data:*

Accommodating the sheer scope of data and creating newer domains in the organization are a challenge, and this can be addressed by implementing a high-performance computing environment and advanced data storage devices like hybrid

storage device which features hard disk drives (HDD) and solid-state drives (SSD); possesses better performance levels with reduced latency, high reliability, and quick access to the data; and therefore helps accumulate large datasets from all the sources. Another way of addressing this challenge can be through discovery of common operational methodologies between the domains for integrating the query operations which stands as a better environment to address the challenges for large data entities.

(ii) *Data inconsistency:*

Data inconsistency refers to the imbalances in data types, structures, and levels. Although the structured data provides the scope for query operations through relational approach so that the data can be analyzed and used by the organization, unstructured data takes a lead always in larger data entities, and this comes as a challenge for organizations. Addressing the data inconsistency can be achieved using the tag and sort methods which allow searching the data using keywords. The new big data tool Hadoop provides the solution for modulating and converting the data through MapReduce and Yarn. Although Hive in Hadoop doesn't support the online transactions, they can be implemented for file conversions and batch processing.

(iii) *Query optimization:*

In real-time data integration, the large data entities require the query optimization at microlevels which could involve mapping components to the existing or a new schema which impacts the existing structures. To address this challenge, the number of queries can be reduced by implementing the joins, strings, and grouping functions. Also the query operations are performed on individual data threads which can reduce the latency and responsiveness. Using the distributed joins like merge, hash, and sort can be an alternative in this scenario but requires more resources. Implementing the grouping, aggregation, and joins can be the best approach to address this challenge.

(iv) *Inadequate resources and implementing support system:*

Lack of resources haunts every organization at certain point, and this has the direct impact on the project. Limited or inadequate resources for creating data integration jobs, lack of skilled labor that don't specialize in data integration, and costs incurred during the implementation of data integration tools can be some of the challenges faced by organizations in real time. This challenge can be addressed by constant resource monitoring within the organization, and limiting the standards to an extent can save the organizations from bankruptcy. Human resources play a major role in every organization, and this could pick the right professionals for the right task in a timely manner for the projects and tasks at hand.

There is a need to establish a support system for updating requirements and error handling, and reporting is required when organizations perform various data integration jobs within the domains and externally. This can be an additional cost for the organizations as setting up a training module to train the professionals and direct them toward understanding the business expectations and deploy them in a fully equipped environment. This can be termed as a good investment as every organization would implement advancements in a timely manner to stick with the growing market trends. Support system for handling errors could fetch them the reviews to analyze the negative feedback and modify the architecture as per the reviews and update the newer versions with better functionalities.

(v) *Scalability:*

Organizations could face big time challenge in maintaining the data accumulated from number of years of their service. This data is stored and maintained using the traditional file systems or other methodologies as per their environment. In this scenario, often the scalability issues arise when the new data from multiple resources is integrated with data from legacy

systems. Changes made by the data scientists and architects could impact the functioning of legacy systems as it has to go through many updates to match the standards and requirements of new technologies to perform a successful data integration. In recent times, mainframe stands as one of the best example for legacy system. For a better data operation environment and rapid access to the data, Hadoop has been implemented by organizations to handle the batch processing unit. This follows a typical ETL (Extract, Transform, and Load) approach to extract the data from number of resources and load them into Hadoop environment for the batch processing.

Some of the common data integration tools which have been in use are Talend, CloverETL, and KARMA. Both the data integration tools have their own significance individually for providing the best data integration solutions for the business.

Real-Time Scenarios for Data Integration

In the recent times, Talend was used as the main base for data integration by Groupon, one of the leading deal-of-the-day website which offers discounted gift certificates, to be used at local shopping stores. For integrating the data from sources, Groupon relied on "Talend." Talend is an open-source data integration tool which is used to integrate data from numerous resources. When Groupon was a startup, they relied on an open source for more gains rather than using a licensed tool which involves more cost for licensing. Since Groupon is a public traded company now, they would have to process 1 TB of data per day, which come from various sources.

There is another case study where a telephone company was facing issues with phone invoices in different formats which were not suitable for electronic processing and therefore involved the manual evaluation of phone bills. This consumed a lot of time and resources for the company. The CloverETL data integration tool was the solution

for the issue, and the inputs given were itemized phone bills, company's employee database, and customer contact database. The data integration process involved consolidated call data records, report phone expenses in hierarchy, and analysis of phone calls and its patterns. This helped organization cut down the costs incurred by 37% yearly.

Conclusion

On a whole, the data integration in current IT world is on demand with the increasing number of data and covers complete aspects of data solutions with the usage of data integration tools. Data scientists are still finding solutions for a simplified data integration with an efficient automated storage systems and visualization methods which could turn out complex in terms of big data. Development of newer data integration solutions in the near future could help address the big data integration challenges. An efficient data integration tool is yet to conquer the market, and evolution of these tools can help organizations handle the data integration in a much more simplified way.

Further Reading

- Big Data Integration. <http://www.ibmbigdatahub.com/video/big-data-integration-what-it-and-why-you-need-it>.
- Clover ETL. <http://www.cloveretl.com/resources/case-studies/data-integration>.
- Data Integration tool. <http://blog.pentaho.com/2011/07/15/facebook-and-pentaho-data-integration/>.
- IBM. How does data integration help your organization? <http://www-01.ibm.com/software/data/integration/>.
- Lenzerini, M. (2002). Data integration: A theoretical perspective, In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 233–246), New York, NY, USA.
- Talend. <https://www.talend.com/customers/customer-reference/groupon-builds-on-talend-enterprise-data-integration>.
- Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP(DOLAP '02)* (pp. 14–21). New York: ACM.

Data Integrity

Patrick Juola

Department of Mathematics and Computer Science, McAnulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Data integrity, along with confidentiality and availability, is one of the three fundamental aspects of data security. Integrity is about ensuring that data is and remains reliable, and that it has not been tampered with or altered erroneously. Hardware or software failures, human mistakes, and malicious actors can all be threats to integrity. *Data integrity* refers specifically to the integrity of the data stored in a system. This can be a particularly critical issue when dealing with big data due to the volume and variety of data stored and processed.

Data integrity deals with questions such as “trust” and “fitness for use” (Lagoze 2014). Even when data has been correctly gathered, stored, and processed, issues of representativeness and data quality can render conclusions unreliable (Lazer et al. 2014).

Data integrity can also be affected by archival considerations. Many big data projects rely on third-party data collection and storage. Text collections such as Wikipedia, Project Gutenberg, and HathiTrust have been used for many language-based big data projects. However, the data in these projects changes over time, as the collections are edited, expanded, corrected, and generally curated. Even relatively harmless fixes such as correcting optical character recognition or optical character reader (OCR) errors can have an effect farther down the processing pipeline; major changes (such as adding documents newly entered into the public domain every year) will cause correspondingly large changes downstream. However, it may not be practical for an organization to archive its own copy of a large database to freeze and preserve it.

Big data technology can create its own data integrity issues. Many databases are too large to

store on a single machine, and even when single-point storage is possible, considerations such as accessibility and performance can lead engineers to use distributed storage solutions such as HBase or Apache Cassandra (Prasad and Agarwal 2016). More machines, in turn, mean more chances for hardware failure, and duplicating data blocks means more chances for copies to become out of sync with each other (which one of the conflicting entries is correct?).

When using a cloud environment (Zhou et al. 2018), these issues are magnified because the data consumer no longer has control of the storage hardware or environment. Cloud storage systems may abuse data management rights and, more importantly, may not provide the desired level of protection against security threats generally (including threats to integrity).

In general, any computing system, even a small, single-system app, should provide integrity, confidentiality, and availability. However, the challenge of providing and confirming data integrity is much harder with projects on the scale of big data.

Further Reading

- Lagoze, C. (2014). Big data, data integrity, and the fracturing of the control zone. *Big Data & Society*, 1(2), 2053951714558281. <https://journals.sagepub.com/doi/abs/10.1177/2053951714558281>.
- Lazer, D., Kennedy, R., & King, G. (2014). The parable of Google flu: Traps in big data analysis. *Science*, 343(6176), 1203–1205.
- Prasad, B. R., & Agarwal, S. (2016). Comparative study of big data computing and storage tools. *International Journal of Database Theory and Application*, 9(1), 45–66.
- Zhou, L., Fu, A., Yu, S., Su, M., & Kuang, B. (2018). Data integrity verification of the outsourced big data in the cloud environment: A survey. *Journal of Network and Computer Applications*, 122, 1–15.

Data Journalism

► [Media](#)

Data Lake

Christoph Quix^{1,2}, Sandra Geisler¹ and Rihan Hai³

¹Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

²Hochschule Niederrhein University of Applied Sciences, Krefeld, Germany

³RWTH Aachen University, Aachen, Germany

Overview

Data lakes (DL) have been proposed as a new concept for centralized data repositories. In contrast to data warehouses (DW), which usually require a complex and fine-tuned Extract-Transform-Load (ETL) process, DLs use a simpler model which just aims at loading the complete source data in its raw format into the DL. While a more complex ETL process with data transformation and aggregation increases the data quality, it might also come with some information loss as irregular or unstructured data not fitting into the integrated DW schema will not be loaded into the DW. Moreover, some data silos might not get connected to integrated data repositories at all due to the complexity of the data integration process. DLs address these problems: they should provide access to the source data in its original format without requiring an elaborated ETL process to ingest the data into the lake.

Key Research Findings

Architecture

Since the idea of a DL has been described first in a blog post by James Dixon (<https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/>), a few DL architectures have been proposed (e.g., Terrizzano et al. 2015; Nargesian et al. 2019). As Hadoop is also able to handle any kind of data in its distributed file system, many people think that “Hadoop” is the complete answer to the question how a DL

should be implemented. Of course, Hadoop is good at managing the huge amount of data with its distributed and scalable file system, but it does not provide detailed metadata management which is required for a DL. For example, the DL architecture presented in Boci and Thistlethwaite (2015) shows that a DL system is a complex eco-system of several components and that Hadoop provides only a part of the required functionality.

More recent articles about data lakes (e.g., Mathis (2017)) mention common functional components of a data lake architecture for ingesting, storing, transforming, and using data. These components are sketched in the architecture of Fig. 1 (Jarke and Quix (2017)).

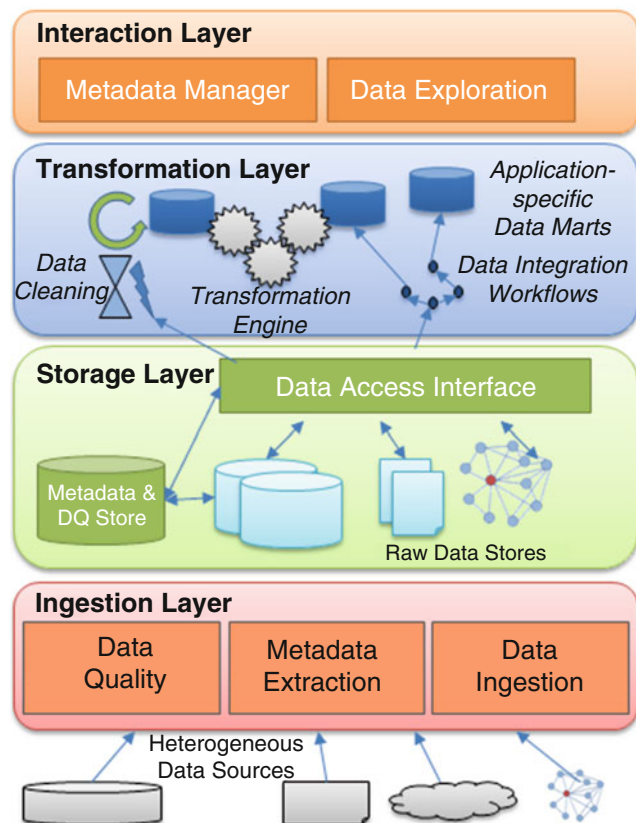
The architecture is separated into four layers: the Ingestion Layer, the Storage Layer, the Transformation Layer, and the Interaction Layer.

Ingestion Layer

One of the key features of the DL concept is the minimal effort to ingest and load data into the DL. The components for data ingestion and metadata extraction should be able to extract data and metadata from the data sources automatically as far as possible, for example, by using methods to extract a schema from JSON or XML. In addition to the metadata, the raw data needs to be also ingested into the DL. According to the idea, that the raw data is kept in its original format, this is more like a “copy” operation and thereby certainly less complex than an ETL process in DWs. Nevertheless, the data needs to be put into the storage layer of the DL, which might imply some syntactical transformation.

Data governance and data quality (DQ) management are important in DLs to avoid data swamps. The Data Quality component should make sure that

Data Lake, Fig. 1 Data lake architecture



the ingested data fulfills minimum data quality requirements. For example, if a source with information about genes is considered, it should also provide an identifier of the genes in one of the common formats (e.g., from the Gene Ontology, <http://geneontology.org>) instead of using proprietary IDs that cannot be mapped to other sources.

Storage Layer

The main components in the storage layer are the metadata repository and the repositories for raw data. The *Metadata Repository* stores all the metadata of the DL which has been partially collected automatically in the ingestion layer or will be later added manually during the curation or usage of the DL.

The raw data repositories are the core of the DL in terms of data volume. As the ingestion layer provides the data in its original format, different storage systems for relational, graph, XML, or JSON data have to be provided. Moreover, the storage of files using proprietary formats should be supported. Hadoop seems to be a good candidate as a basic platform for the storage layer, but it needs to be complemented with components to support the data fidelity, such as Apache Spark. In order to provide a uniform way to the user to query and access the data, the hybrid data storage infrastructure should be hidden by a uniform *data access interface*. This data access interface should provide a query language and a data model, which have sufficient expressive power to enable complex queries and represent the complex data structures that are managed by the DL. Current systems such as Apache Spark and HBase offer this kind of functionality using a variant of SQL as query language and data model.

Transformation Layer

To transform the raw data into a desired target structure, the DL needs to offer a data transformation engine in which operations for data cleaning, data transformation, and data integration can be realized in a scalable way. In contrast to a data warehouse, which aims at providing one integrated schema for all data sources, a DL should support the ability to create *application-specific* data marts, which integrate a subset of the raw

data in the storage layer for a concrete application. From a logical point of view, these data marts are rather part of the interaction layer, as the data marts will be created by the users during the interaction with the DL. On the other hand, their data will be stored in one of the systems of the storage layer. In addition, data marts can be more application-independent if they contain a general-purpose dataset which has been defined by a data scientist. Such a dataset might be useful in many information requests of the users.

Interaction Layer

The top layer focuses at the interaction of the users with the DL. The users will have to access the metadata to see what kind of data is available, and can then explore the data. Thus, there needs to be a close relationship between the data *exploration* and the *metadata manager* components. On the other hand, the metadata generated during data exploration (e.g., semantic annotations, discovered relationships) should be inserted into the metadata store. The interaction layer should also provide to the user functionalities to work with the data, including visualization, annotation, selection, and filtering of data, and basic analytical methods. More complex analytics involving machine learning and data mining is in our view not part of the core of a DL system, but certainly a very useful scenario for the data in the lake.

Data Lake Implementations

Usually, a DL is not an out-of-the-box ready-to-use system. The above described layers have to be assembled and configured one by one according to the organization's needs and business use cases. This is tedious and time consuming, but also offers a lot of flexibility to use preferred tools. Slowly, implementations which offer some of the previously described features of DLs in a bundle are evolving. The Microsoft Azure Data Lake (<https://azure.microsoft.com/solutions/data-lake>) offers a cloud implementation especially for the storage layer building on HDFS enabling a hierarchical file system structure. Another implementation covering a wide range of the above mentioned features is offered by the open-source data lake management platform Kylo (<https://kylo.io/>).

Kylo is built on Hadoop, Spark, and Hive and offers out-of-the-box wrappers for streams as well as for batch data source ingestion. It provides meta-data management (e.g., schema extraction), data governance, and data quality features, such as data profiling, on the ingestion and storage layer. Data transformation based on Spark and a common search interface are also integrated into the platform. Another advanced but commercial platform is the Zaloni Data Platform (<https://www.zaloni.com/platform>).

Future Directions for Research

Lazy and Pay-as-You-Go Concepts

A basic idea of the DL concept is to consume as little upfront effort as possible and to spend additional work during the interaction with users, for example, schemas, mappings, and indexes are created while the users are working with the DL. This has been referred to as lazy (e.g., in the context of loading database (Karæz et al. 2013)) and pay-as-you-go techniques (e.g., in data integration (Sarma et al. 2008)). All workflows in a DL system have to be verified, whether the deferred or incremental computation of the results is applicable in that context. For example, meta-data extraction can be done first in a shallow manner by extracting only the basic metadata; only detailed data of the source is required, a more detailed extraction method will be applied. A challenge for the application of these “pay-as-you-go” methods is to make them really “incremental,” for example, the system must be able to detect the changes and avoid a complete recomputation of the derived elements.

Schema-on-Read and Evolution

DLs also provide access to un- or semi-structured data for which a schema was not explicitly given during the ingestion phase. Schema-on-read means that schemas are only created when the data is accessed, which is inline with the “lazy” concept described in the previous section. The shallow extraction of metadata might also lead to changes at the metadata level, as schemas are

being refined if more details of the data sources are known. Also, if a new data source is added to the DL system, or an existing one is updated, some integrated schemas might have to be updated as well, which leads to the problem of schema evolution (Curino et al. 2013).

Another challenge to be addressed for schema evolution is the heterogeneity of the schemas and the frequency of the changes. While data warehouses have a relational schema which is usually not updated very often, DLs are more agile systems in which data and metadata can be updated very frequently. The existing methods for schema evolution have to be adapted to deal with the frequency and heterogeneity of schema changes in a big data environment (Hartung et al. 2011).

Mapping Management

Mapping management is closely related to the schema evolution challenge. Mappings state how data should be processed on the transformation layer, that is, to transform the data from its raw format as provided by the storage layer to a target data structure for a specific information requirement. Although heterogeneity of data and models has been considered and generic languages for models and mappings have been proposed (Kensche et al. 2009), the definition and creation of mappings in a schema-less world has not received much attention, yet. The raw data in DLs is less structured and schema information is not explicitly available. Thus, in this context methods for data profiling or data wrangling have to be combined with schema extraction, schema matching, and relatable dataset discovery (Alserafi et al. 2017; Hai et al. 2019).

Query Rewriting and Optimization

In the data access interface of the storage layer, there is a trade-off between expressive power and complexity of the rewriting procedure as the complexity of query rewriting depends to a large degree on the choice for the mapping and query language. However, query rewriting should not only consider completeness and correctness, but also the costs for executing the rewritten query should be taken into account.

Thus, the methods for query rewriting and query optimization require a tighter integration (Gottlob et al. 2014). It is also an open question whether there is a need for an intermediate language in which data and queries are translated to do the integrated query evaluation over the heterogeneous storage system, or whether it is more efficient to use some of the existing data representations. Furthermore, given the growing adoption of declarative languages in big data systems, query processing and optimization techniques from the classical database systems could be applied as well in DL systems.

Data Governance and Data Quality

Since the output of a DL should be useful knowledge for the users, it is important to prevent a DL becoming a *data swamp*. There are conflicting goals: on the one hand, any kind of data source should be accepted for the DL, and no data cleaning and transformation should be necessary before the source is ingested into the lake. On the other hand, the data of the lake should have sufficient quality to be useful for some applications. Therefore, it is often mentioned that data governance is required for a DL. First of all, data governance is an organizational challenge, that is, roles have to be identified, stakeholders have to be assigned to roles and responsibilities, and business processes need to be established to organize various aspects around data governance (Otto 2011). Still, data governance needs to be also supported by appropriate techniques and tools. For data quality, as one aspect of data governance, a similar evolution has taken place; the initially abstract methodologies have been complemented in the meantime by specific techniques and tools. Preventive, process-oriented data quality management (in contrast to data cleaning, which is a reactive data quality management) also addresses responsibilities and processes in which data is created in order to achieve a long-term improvement of data quality.

Data Models and Semantics in Data Lakes

While it has been acknowledged that metadata management is an important aspect in DLs,

there are only few works on the modeling of data and metadata in a DL. Data vault is a dimensional modeling technique frequently applied in DW projects; in Giebler et al. (2019), this modeling technique is applied to DLs and compared with other techniques. Because of the fragmentation of the data in many different tables, querying is expensive due to many join operations. Also, the mapping of DLs to semantic models has been considered in Endris et al. (2019). They propose a framework that maps heterogeneous sources to a unified RDF graph and thereby allows federated query processing. Still, there is a need for more sophisticated metadata models and data modeling techniques for DLs to provide more guidance in managing a DL.

Cross-References

- Big Data Quality
- Data Fusion
- Data Integration
- Data Quality Management
- Data Repository
- Metadata

Further Reading

- Alserafi, A., Calders, T., Abelló, A., & Romero, O. (2017). Ds-prox: Dataset proximity mining for governing the data lake. In C. Beecks, F. Borutta, P. Kröger, & T. Seidl (Eds.), *Similarity search and applications -10th international conference, SISAP 2017, Munich, Germany, October 4–6, 2017, proceedings* (Vol. 10609, pp. 284–299). Springer. <https://doi.org/10.1007/978-3-319-68474-120>.
- Boci, E., & Thistlethwaite, S. (2015). A novel big data architecture in support of ads-b data analytic. In *Proceedings of the integrated communication, navigation, and surveillance conference (icns)* (pp. C1-1–C1-8). <https://doi.org/10.1109/ICNSURV.2015.7121218>.
- Curino, C., Moon, H. J., Deutsch, A., & Zaniolo, C. (2013). Automating the database schema evolution process. *VLDB Journal*, 22(1), 73–98.
- Endris, K. M., Rohde, P. D., Vidal, M., & Auer, S. (2019). Ontario: Federated query processing against a semantic data lake. In *Proceedings of 30th international*

conference on database and expert systems applications (dexa) (Vol. 11706, pp. 379–395). Springer. Retrieved from https://doi.org/10.1007/978-3-030-27615-7_29.

- Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Modeling data lakes with data vault: Practical experiences, assessment, and lessons learned. In *Proceedings of the international conference on conceptual modeling (er)*. (to appear).
- Gottlob, G., Orsi, G., & Pieris, A. (2014). Query rewriting and optimization for ontological databases. *ACM Transactions on Database Systems*, 39(3), 25:1–25:46. Retrieved from <https://doi.org/10.1145/2638546>.
- Hai, R., Quix, C., & Wang, D. (2019). Relaxed functional dependency discovery in heterogeneous data lakes. In *Proceeding of the international conference on conceptual modeling (er)*. (to appear).
- Hartung, M., Terwilliger, J. F., & Rahm, E. (2011). Recent advances in schema and ontology evolution. In Z. Bellahsene, A. Bonifati, & E. Rahm (Eds.), *Schema matching and mapping* (pp. 149–190). Springer Berlin/Heidelberg. Retrieved from <https://doi.org/10.1007/978-3-642-16518-4>.
- Jarke, M., & Quix, C. (2017). On warehouses, lakes, and spaces: The changing role of conceptual modeling for data integration. In J. Cabot, C. Gómez, O. Pastor, M. Sancho, & E. Teniente (Eds.), *Conceptual modeling perspectives* (pp. 231–245). Springer. <https://doi.org/10.1007/978-3-319-67271-716>.
- Karæz, Y., Ivanova, M., Zhang, Y., Manegold, S., & Kersten, M. L. (2013). Lazy ETL in action: ETL technology dates scientific data. *PVLDB*, 6(12), 1286–1289. Retrieved from <http://www.vldb.org/pvldb/vol6/pl1286-kargin.pdf>.
- Kensche, D., Quix, C., Li, X., Li, Y., & Jarke, M. (2009). Generic schema mappings for composition and query answering. *Data & Knowledge Engineering*, 68(7), 599–621. <https://doi.org/10.1016/j.datak.2009.02.006>.
- Mathis, C. (2017). Data lakes. *Datenbank-Spektrum*, 17(3), 289–293. <https://doi.org/10.1007/s13222-017-0272-7>.
- Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: Challenges and opportunities. *PVLDB*, 12(12), 1986–1989. Retrieved from <http://www.vldb.org/pvldb/vol12/p1986-nargesian.pdf>.
- Otto, B. (2011). Data governance. *Business & Information Systems Engineering*, 3(4), 241–244. <https://doi.org/10.1007/s12599-011-0162-8>.
- Sarma, A. D., Dong, X., & Halevy, A. Y. (2008). Bootstrapping pay-as-you-go data integration systems. In J. T.-L. Wang (Ed.), *Proceedings of ACM SIGMOD international conference on management of data* (pp. 861–874). Vancouver: ACM Press.
- Terrizzano, I., Schwarz, P. M., Roth, M., & Colino, J. E. (2015). Data wrangling: The challenging journey from the wild to the lake. In *7th biennial conference on innovative data systems (cidr)*. Retrieved from http://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper2.pdf.

Data Management and Artificial Intelligence (AI)

Alan R. Shark

Public Technology Institute, Washington, DC, USA

Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

The numbers are staggering as big data keeps getting bigger. We know that over 300 hours of YouTube videos are downloaded every minute of every day. Google alone processes more than 40,000 searches every second, and when other search engines are added-combined, they account for some 5 billion searches a day worldwide (<https://merchdope.com/youtube-stats/>). When we look at all the text messages and posted and shared pictures, let alone emails and other forms of digital communications, we create no less than 2.5 quintillion bytes of data each day (<https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#3bdc425b60ba>). These statistics were published prior to the COVID-19 pandemic where the growth of video feeds and storage have grown hundreds of percent over a short period of time.

Data curation and data mining have become a growing specialty. Data has been used to help understand the opioid crisis with visualized map planning and has been used to better understand the outbreak of the COVID-19 pandemic. As data continues to accumulate, locating and analyzing data in a timely manner become a never-ending challenge. Many are turning to artificial intelligence (AI) to assist. The potential to harness data with AI holds enormous promise. But there are some roadblocks to navigate around, and one must take a deeper dive to better understand the relationship between the two.

AI, at least how it is applied today, is not as new as some would believe. Machine learning (ML) has been around for some 50+ years and can be defined as the scientific study of algorithms and statistical models that computer

systems use to carry out tasks without explicit instructions, such as by using pattern recognition and inference (https://en.wikipedia.org/wiki/Machine_learning). Spam and e-mail filters are good examples of ML where algorithms are constantly being updated to detect either which emails to be placed in which folders or which emails should be considered SPAM. Note, even the best filters still allow for humans to check to make sure that any email considered SPAM is just that – and there are times when things are judged incorrectly.

ML is a large departure from the beginnings of machine programing where computers relied totally on programs and instructions. Today, through ML, machines can be programed to seek out patterns and are the cornerstone for predictive analytics (https://en.wikipedia.org/wiki/Machine_learning). Machine learning can be viewed as a unique subfield of artificial intelligence in which algorithms learn to fulfill tasks. AI in practice is being developed to mimic human behavior. AI can best be described as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation (<https://botanalytics.co/blog/2017/08/18/machine-learning-artificial-intelligence/>).

So, to harness the power of data, AI can be used to search trillions upon trillions of pieces of data in seconds or less in search of a pattern, anomaly, statistical probability, and simply anything a human might perform in seconds versus years or more. Several futurists believe in time that AI will move from machine intelligence to machine consciousness. A popular example of machine learning would be our growing reliance on such devices such as Alexa, Siri, and talking into our TV remotes. An example of machine consciousness might be talking to a robot who expresses human emotions and can appear to both think, feel, and reason.

AI has advanced in the past 10 years because of six key factors; they are:

1. Advancements in complex algorithms
2. Dramatic increase in speed and computing power
3. Ability to digest data from various sources

4. Ability to store and retrieve massive amounts of data
5. Ability to “self-learn”
6. Advancements in artificial speech and recognition

This author, through studying the practical applications of AI, has reached the conclusion (at least at this writing) that a more accurate definition would be “the theory and development of computer systems able to supplement human decision making, planning and forecasting based on abundant sources of quality data.” Thus, what we have today is AI as *augmented intelligence*, assisting humans in searching for meaningful answers to contemporary problems and helping to make decisions based on data.

On February 11, 2019, President Trump signed Executive Order 13859 announcing the *American AI Initiative – the United States’ National Strategy on Artificial Intelligence* (<https://www.whitehouse.gov/ai/>). Aside from promoting AI in the government workspace through collaboration with industry and academia, there was a clear recognition that data bias and ethics need to be addressed as AI applications advance (<https://www.whitehouse.gov/ai/ai-american-innovation/>). Many have warned of the potential dangers of AI if ethics and bias are not adequately tackled. Can AI in collecting data through articles and papers distinguish between peer-reviewed studies versus opinions that may reflect poor or lack of any scientific proof – or worse – ignorant summations, conspiracy theories, or racist leanings? (<https://www.napawash.org/studies/academy-studies/ai-and-its-impact-on-public-administration>).

But the USA was not the first among the most developed nations to develop its AI initiative. The European Union developed its trustworthy AI initiative in early 2018 and articulated seven basic principles which are (<https://ec.europa.eu/digital-single-market/en/artificial-intelligence>):

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, nondiscrimination, and fairness

6. Societal and environmental well-being
7. Accountability

And most observers believe China has taken the lead in AI development, and the rest of the Western worlds is trying to catch up.

Regardless of which nation is developing AI strategies and applications, there is one universal truth – when it comes to data, we have always been “taught garbage in equals garbage out.” If not properly trained or applied, AI can produce harmful results especially if there is no mechanism in place to validate data and to audit what data and streams entered a recommendation or decision.

Often, we learn more from our mistakes than from our successes. A case in point involves IBM’s Watson. Watson has been IBM’s flagship platform for AI and was popularized in winning games of jeopardy on TV as well as beating a leading international chess player at chess. But one would learn that Watson’s medical applications did not live up to its promise. IBM stated its goal to the public that it hoped to create AI doctor. But that never materialized, so they further refined Watson. However, while medical Watson learned rather quickly how to scan articles about clinical studies and determine basic outcomes, it proved impossible to teach Watson to read the articles the way a doctor would. Later they would learn more on how doctors absorb information they read from journal articles. They found that doctors often would look for other pieces of information that may not have been the main point of the article itself. This was not an anticipated outcome (<https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care>).

For years IBM programmers worked on developing a system that could assist doctors in diagnosing cancer as an augmented decision-maker, but real-life experience taught them something else. Due to weaknesses in quality control later found in the way data was collected and cataloged, results were often misleading and could have ultimately led to a patient’s death. It had to be pulled from the market – at least for now.

Data management and data ingestion are essential components of AI. Data is often collected without any idea of how it might be used at a later time; therefore it is imperative that data be tagged in ways that make it easier to comprehend (human or machine) any limitations or bias. Data quality and data management have never been more important.

Data management requires a renewed focus on policies and procedures. While data management is applied in many ways depending on any institution, one may look to the US federal government for some examples and guidance of how they plan to modernize and standardize data management. In March 2018, as part of the President’s Management Agenda (PMA), the administration established a cross-agency priority (CAP) goal focused on leveraging data as a strategic asset to establish best practices for how agencies manage and use data. As part of this CAP goal, the first-ever enterprise-wide Federal Data Strategy (FDS) was developed to establish standards, interoperability, and skills consistency across agencies (<https://strategy.data.gov/action-plan/>).

Of the 20 action steps highlighted, there are 2 that directly apply to this discussion, action steps 8 and 19. Step 8 to aims “Improve Data and Model Resources for AI Research and Development” is by its title directly related to AI. Step #8 directly ties back to the 2019 Executive Order on Maintaining American Leadership in Artificial Intelligence (<https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/>).

The order includes an objective to “Enhance access to high-quality and fully traceable federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections consistent with applicable laws and policies.” The implementation guidance provides support to agencies in:

- Prioritizing the data assets and models under their purview for discovery, access, and enhancement

- Assessing the level of effort needed to make necessary improvements in data sets and models, against available resources
- Developing justifications for additional resources

Action step 19 sets out to “Develop Data Quality Measuring and Reporting Guidance.” As pointed out earlier, data quality is essential to advance AI. One of the key objectives is to identify best practices for measuring and reporting on the quality of data outputs created from multiple sources or from secondary use of data assets (<https://strategy.data.gov/action-plan/#action-19-develop-data-quality-measuring-and-reporting-guidance>). Clearly data management is essential to the success and meaningful application of AI.

There is no doubt that AI will revolutionize the way we collect and interpret data into useful and actionable information. Despite some early disappointments, AI continues to hold great promise if ingesting trillions of pieces of data can lead to incredible medical and social science outcomes. It can be used in understanding and solving complex issues regarding public policy and public health and safety. AI holds the promise of making rational determinations based on big data – and pulling from an array of sources from among structured and unstructured data sets.

With all the attention paid to AI’s potential regarding data management and AI, there are also many related concerns beyond bias and ethics, and that is privacy. The European Union gained much attention in 2016 when it passed its landmark GDPR or General Data Privacy Act. This act articulated seven basic principles – all aimed at protecting and providing redress for inaccuracies and violations and how data is viewed and how long it is stored. To help clarify and manage the GDPR, the EU established *The European Data Protection Board (EDPB)* in 2018 (https://edpb.europa.eu/about-edpb/about-edpb_en). It provides general guidance (including guidelines, recommendations, and best practice) to clarify the GDPR and advises the European Commission on data protection issues and any proposed new EU legislation of particular importance for the protection of personal data and

encourages national data protection authorities to work together and share information and best practices with one another.

The US government has thus far resisted a national approach to data privacy which prompted the State of California to pass its own privacy act called *The California Consumer Privacy Act (CCPA)* (<https://www.oag.ca.gov/privacy/ccpa>). As data management becomes more sophisticated through technology and governance, the public has demonstrated growing concern over privacy rights. When you add AI as the great multiplier of data, a growing number of citizen advocates are seeking ways to protect their personal data and reputations and seek remedies and procedures to correct mistakes or to limit their exposure.

As Western nations struggle to protect privacy and maintain a healthy balance between personal information and the need to provide better economic, public health, and safety outcomes, AI continues to advance in ways that continue to defy our imaginations. Any sound data management plan must certainly address the issue of privacy.

For AI to reach its potential in the years ahead, data management will always play a significant supporting role. Conversely, poor data management will yield disappointing results that could ultimately lead to less than optimal outcomes, and worse could lead to the loss of lives. Sound data management is what feeds AI, and AI requires nothing less than quality and verified intake.

Data Mining

Gordon Alley-Young

Department of Communications and Performing Arts, Kingsborough Community College, City University of New York, New York, NY, USA

Synonyms

[Anomaly detection](#); [Association analysis](#); [Cell phone data](#); [Cluster analysis](#); [Data brokers](#); [Data mining algorithms](#); [Data warehouse](#); [Education](#);

Facebook; National Security Administration (NSA); Online analytical processing; Regression

Introduction

Data mining (DM), also called knowledge discovery in data (KDD), examines vast/intricate stores of citizen, consumer, and user data to find patterns, correlations, connections and/or variations in order to benefit organizations. DM serves to either describe what is happening and/or to predict what will happen in the future based on current data. Information discovered from DM falls into several categories including: associations, anomalies, regressions, classifications, and clusters. DM uses algorithms and software programs to analyze data collections. The size and complexity of a data collection also called a data warehouse (DWH) will determine how sophisticated the DM system will need to be. The DM industry was estimated to be worth 50 billion dollars in 2017, and aspects of the DM industry that help companies eliminate waste and capitalize on future business trends are said to potentially increase an organization's profitability threefold. Social media organizations profit from DM services that coordinate advertising and marketing to its users. DM has raised the concerns of those who fear that DM violates their privacy.

History of DM

DM was coined during the 1960s but its roots begin in the 1950s with innovations in the areas of artificial intelligence (AI) and computer learning (CL). Initially DM was two separate functions: one that focused on retrieving information and another that dealt with database processing. From the 1970s to the 1990s, computer storage capacity and the development of computer programming languages and various algorithms advanced the field, and by the 1990s Knowledge Discovery in Databases (KDD) was actively being used. Falling data storage costs and rising computer processing rates during this decade meant that many businesses now used KDD or

DM to manage all aspects of their customer relations.

The 1990s saw the emergence of online analytical processing (OLAP) or computer processing that quickly and easily chooses and analyzes multidimensional data from various perspectives. OLAP and DM are both similar and different, and thus can be used together. OLAP can summarize data, distribute costs, compile and analyze data over time periods (time series analysis), and do what-if analysis (i.e., what if we change this value, how will it change our profit or loss?). Unlike DM, OLAP systems are not CL systems in that they cannot find patterns that were not identified previously. DM can do this independent pattern recognition, also called machine learning (ML), which is an outgrowth of 1950s AI inquiry. DM can inductively learn or make a general rule once it observes several instances. Once DM discovers a new pattern and identifies the specific transaction data for this pattern, then OLAP can be used to track this new pattern over time. In this way OLAP would not require a huge data warehouse (DWH) like DM does because OLAP has been configured to only examine certain parts of the transaction data.

Types of DM

DM's capacity for intuitive learning falls into five categories: associations, anomalies, regressions, classifications, and clusters. Association learning (AL) or associations is DM that recommends products, advertisements, coupons and/or promotional e-mails/mailings based on an online customer's purchasing profile or a point of sale (POS) customer's data from scanned customer loyalty cards, customer surveys, and warranty purchases/registrations. AL allows retailers to create new products, design stores/websites, and stock their products accordingly. AL uses association learning algorithms (ALA) to do what some have called market basket analysis. Instead of finding associations, Anomaly detection (AD), also called anomalies, outlier detection (OD), or novelty detection (ND) uses anomaly detection algorithms (ADA) to find phenomena outside of

a pattern. Credit card companies use AD to identify possible fraudulent charges and governmental tax agencies like the Internal Revenue Service (IRS) in the USA, Canada Revenue Agency (CRA) in Canada, and Her Majesty's Revenue and Customs (HMRC) in the UK use it to find tax fraud.

In 2014 Facebook bought messaging service WhatsApp in order, analysts argue, to expand their DM capabilities and increase profits. Founded by two former Yahoo employees, WhatsApp allows users to send text, photo, and voice messages to and from any smartphone without a special setup for a small annual fee. Since its release in 2009, WhatsApp has gained 500 million users and by spring 2014 the company announced that its users set a new record by sending 64 billion messages in a 24 h period. Owning WhatsApp gives Facebook access to users' private messaging data for purposes of AL and the better marketing of products to consumers. This is because consumers' private data may be more accurate, for marketing purposes, than the public information they might share on open platforms like Facebook. This is especially important in countries like India, with an estimated 50 million WhatsApp users (100 million Facebook users) whose primary communication device is mobile.

Regression analysis (RA) makes predictions for future activity based on current data. For example, a computer application company might anticipate what new features and applications would interest their users based on their browsing histories; this can shape the principles they use when designing future products. RA is done using regression analysis algorithms (RAA). Much of DM is predictive and examines consumers' demographics and characteristics to predict what they are likely to buy in the future. Predictions have probability (i.e., Is this prediction likely to be true?) and confidence (i.e., How confident is this prediction?).

Classification analysis (CA) is DM that operates by breaking data down into classes (categories) and then breaking down new examples into these classes (categories). CA works by creating and applying rules that solve categorization problems. For example, an Internet provider

could employ a junk mail/spam filter to prevent their customers from being bombarded with junk e-mails so that the filter uses CA over time to learn to recognize (e.g., from word patterns) what is spam and what is not. Hypothetically, companies could also use CA to help them to design marketing e-mails to avoid junk mail/spam filters. This form of DM uses classification analysis algorithms (CAA).

The final type of DM is called cluster detection (CD) or clusters and is used for finding clusters of data (e.g., individuals) that are homogenous and distinct from other clusters. Segmentation algorithms (SA) create clusters of data that share properties. CD has been used by the retail industry to find different clusters of consumers among their customers. For example, customers could be clustered based on their purchasing activities where a cluster of one time impulse buyers will be distinct from the cluster of consumers who will continually upgrade their purchases or the cluster of consumers who are likely to return their purchases. The National Security Administration (NSA) in the USA and Europol in the European Union (EU) uses CD to locate clusters of possible terrorists and may use CD on telephone data. For example, if a known terrorist calls a citizen in the USA, France, Belgium, or Germany, the contacts of that contacted person creates a data cluster of personal contacts that could be mined to find people who may be engaged in terrorist activity. CD data can be used as evidence to justify potentially obtaining a warrant for wiretapping/lawful interception or for access to other personal records.

Government, Commercial, and Other Applications of DM

Governmental use of DM has proven controversial. For example, in the USA, in June 2013, former government contractor Edward Snowden released a classified 41-slide PowerPoint presentation found while working for the NSA to two journalists. Snowden charged that a NSA DM program, codenamed PRISM, alleged to cost 20 million dollars a year, collected Internet user's photos, stored data, file transfers, e-mails,

chats, videos, and video conferences with the technology industry's participation. These materials allege that Google and Facebook starting in 2009, YouTube starting in 2010, AOL starting in 2011, and Apple starting in 2012 were providing the government with access to users' information. Subsequent to the release of this information, the USA cancelled Snowden's passport and Snowden sought temporary asylum in Russia to escape prosecution for his actions.

At an open House Intelligence Committee Meeting in June of 2013, FBI Deputy Director Sean Joyce claimed that the PRISM program allowed the FBI in 2009 to identify and arrest Mr. Najibullah Zazi, an airport shuttle driver from outside Denver, who was subsequently convicted in 2010 of conspiring to suicide bomb the New York City Subway system. Snowden has alleged that low-level government analysts have access to software that searches hundreds of databases without proper oversight. The White House denied abusing the information it collects and countered that analyzing big data (BD) (i.e., extremely large and complex data sets) was helping to make the US government work better by eliminating \$115 million in fraudulent medical payments and protecting national security all while preserving privacy and civil rights. While not claiming that the USA has abused privacy or civil rights, civil libertarians have spoken out against the potential for misuse that are represented by programs like PRISM.

In addition to governmental use, retail and social media industries DM is used in education and sports organizations. Online education uses DM on data collected during computer instruction, in concert with other data sources, to assess and improve instruction, courses, programs, and educational outcomes by determining when students are most/least engaged, on/off task or experiencing certain emotional states (e.g., frustration). Data collected include students' demographics, learning logs/journals, surveys, and grades. The National Basketball Association (NBA) has used a DM program called Advanced Scout to analyze data including image recordings

of basketball games. Advanced Scout analyzes player movements and outcomes to predict successful playing strategies for the future (i.e., RA) and also to find unusual game outcomes (i.e., like a player's scoring average differing considerably from usual patterns (i.e., AD)). The Union of European Football Associations (UEFA) and its affiliated leagues/clubs similarly collaborate with private DM firms for results and player tracking as well as for developing effective team rosters.

Every transaction or interaction yields data that is captured and stored making the amounts of data unmanageable by human means alone. The DM industry has spawned a lucrative data broker (DB) industry (e.g., selling consumer data). Three of the largest DB's are companies Acxiom, Experian, and Epsilon who do not widely publicize the exact origins of their data or the names of their corporate customers. DB company Acxiom is reportedly the second largest of the three DBs with approximately 23,000 computer servers that process over 50 trillion yearly data transactions. Acxiom is reported to have records on hundreds of millions US citizens with 1,500 pieces of data per consumer (e.g., Internet browser cookies, mobile user profiles). Acxiom partnered with company DataXpand to study consumers in Latin America and Spain as well as US-Hispanics and Acxiom also has operations in Asia and the EU.

Conclusion

It was estimated in 2012 that the world created 2.5 quintillion bytes of data or 2.5×10^{18} bytes (i.e., 2.5 exabytes (EB)) per day. To put this in perspective, in order to store 2.5 quintillion bytes of data, one would need to use over 36 million 64-gigabyte (GB) iPhones as this number of devices would only provide an estimated 2.30400 EB of storage. Of the data produced, it is estimated that over 30% is nonconsumer data (e.g., patient medical records) and just less than 70% is consumer data (e.g., POS data, social media). By 2019, it is estimated that the emerging

markets of Brazil, Russia, India, and China (BRIC economies) will produce over 60% of the world's data. Increases in global data production will increase the demand for DM services and technology.

Cross-References

► [Business Intelligence Analytics](#)

References

- Executive Office of the President. (2014). *Big data: Seizing opportunities, preserving values*. Retrieved from http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.
- Frاند, J. (n.d.). *Data mining: What is data mining?* Retrieved from <http://www.anderson.ucla.edu/faculty/jason.frاند/teacher/technologies/palace/datamining.htm>.
- Furnas, A. (2012). *Everything you wanted to know about data mining but were afraid to ask*. Retrieved from <http://www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388/>.
- Jackson, J. (2002). Data mining: A conceptual overview. *Communications of the Association for Information Systems*, 8, 267–296.
- Oracle. (2016). *What is data mining?* Retrieved from https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDFGCIJ.
- Pansare, P. (2014). *You use Facebook: Or Facebook is using you?* Retrieved from http://epaper.dnaindia.com/story.aspx?id=66809&boxid=30580&ed_date=2014-07-01&ed_code=820040&ed_page=5.
- Pappalardo, J. (2013). NSA data mining: How it works. *Popular Mechanics*, 190(9), 59.
- US Senate Committee on Commerce, Science and Transportation. (2013). *A review of the data broker industry: Collection, use, and sale of consumer data for marketing purposes: Staff report for chairman Rockefeller*. Retrieved from http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577.
- Yettick, H. (2014). Data mining opens window on student engagement. *Education Week*, 33, 23.

Data Monetization

Rhonda Wrzenski¹, R. Bruce Anderson^{2,3} and Corey Koch³

¹Indiana University Southeast, New Albany, IN, USA

²Earth & Environment, Boston University, Boston, MA, USA

³Florida Southern College, Lakeland, FL, USA

Given advances in technology, companies all around the globe are now collecting data in order to better serve their customers, limit the infringement of company competitors, maximize profits, limit expenditures, reduce corporate risk-taking, and maintain productive relationships with business partners. This utilization and monetization of big data has the potential to reshape corporate practices and to create new revenue streams either through the creation of platforms that consumers or other third parties can interface with or through revamped business practices that are guided by data.

At its simplest, data monetization is the process of generating revenue from some source of data. Though recent technological developments have triggered exponential growth of the field, data monetization dates back to the 1800s with the first mail order catalogs.

In order to monetize data, a data supply chain must exist and be utilized for the benefit of corporations or consumers. The data supply chain is composed of three main components: a data creator, a data aggregator, and a data consumer. In the process of data monetization, data must first be created. That created data must then be available for use. The data can then be recognized, linked, amassed, qualified, authenticated, and, finally, traded. In the process, the data must be stored securely to prevent cyber attacks.

An individual data creator must input data. This could be a person utilizing some type of application programming interface or a sensor, database, log, web server, or software algorithm of some sort that routinely collects or generates

Data Mining Algorithms

► [Data Mining](#)

the data. The data source can be compiled over time or instantaneously streamed.

Institutions also generate data on a large scale to include their transactions and computer interactions. This data generation also puts institutions at risk of unauthorized extraction of their generated data.

The value of this data comes from its ability to be researched and analyzed for metadata and outcome information. Outcome information can be simple knowledge from the data. This information can be used to streamline production processes, enhance development, or formulate predictions of future activity. The entire logistics field has come to be based on self-data analysis for the sake of some type of corporate improvement. When this improvement comes to cause an increase in profitability, the data was effectively monetized.

When individuals create data, they are not typically compensated. Legally, the individual has ownership of the data he or she creates. However, many Internet interfaces, such as Google, require a data ownership waiver for users to access their site. This means that Google owns all user data from its site by default. However, without such a waiver, the information belongs to its creator, the individual user. Alas, many users do not recognize that they waive their personal data rights. Moreover, some users simply access the site without an official account, meaning that they have not specifically waived their personal data ownership rights. Nonetheless, Google still aggregates this user data for profit.

Once data is created, it must be aggregated. Data aggregation is the process of compiling information from varying sources. The aggregation of data can be used for many things, from scientific research to commercial advancements. For instance, by aggregating Covid-19 data from various sources like the World Health Organization (WHO), the European Center for Disease Control and Prevention (ECDC), and the Centers for Disease Control and Prevention (CDC), one can better track the cases, death rates, hospitalization status, and recovery of patients around the globe.

After data has been aggregated, it can be sold. However, raw data is much like crude oil: it requires refining. Data is typically processed, abridged, and stored. At this point, the data is known as processed data. From processed data derives data insights, which encompass the fields of data mining, predictive modeling, analytics, and hard data science, which are defined in more detail below. Through these specialized fields, data has its true value; these fields bring about the benefits data has to offer. Once data has been processed in these fields, outcome information is available. Outcome information is the final product of the data refining process and can be used as business intelligence to improve commerce.

Data mining is the term used to describe the practice of analyzing large amounts of data to discern patterns and relationships in the data. From this, decisions can be made that are aided by the information. The compilation of a profile based on this information is known as data profiling. Data mining and profiling produced finished products from the data supply and refinement chain that can be used for a variety of practical applications. These products are not only valuable for traditional commercial applications but also for political and defense applications. For instance, political candidates can use big data to target segments of the voting population with direct mail or advertisements that appeal to their interests or policy preferences.

Because of the volatility and power of data in the modern age, government agencies like the Federal Bureau of Investigation (FBI) and the National Security Agency (NSA) in the United States are regularly involved in data aggregation and consumption itself. For example, the NSA maintains data on call detail records and social media announcements. This information is processed and used for counterterrorism initiatives and to monitor or predict criminal activity. The NSA contracts with various companies, such as Dataminr and Venntel, Inc., to compile data from a multitude of sources.

Predictive modeling is the creation of models based on probability. Most often used to predict

future outcomes, predictive modeling is a product of data refinement that can be applied to nearly any kind of unknown event. One example of how this technique can be used by corporations would be a business using data on consumer purchasing behavior to target those most likely to shop with special digital or in-store coupons.

Data analytics is a broader field that encompasses the examination, refining, transforming, and modeling of data. Data analysis refers to the process of scrutinizing data for useful information.

Data science is the even broader field that involves using statistical methods and algorithms to extract and derive general knowledge from data. For instance, a data scientist can create a dashboard to allow consumers or clients to interface with the data, to track trends, and to visualize the information.

Some industries operate under federal regulation in the United States that prohibits the free sharing of consumer personal information. The Health Information Technology for Economic and Clinical Health Act (HITECH) and the Health Insurance Portability and Accountability Act (HIPAA) are two laws in the healthcare industry within the United States that limit the ready transfer of personal information. Nonetheless, healthcare corporations recognize the value of data and operate within these federal confines to offer somewhat synchronized care systems. This data sharing contains the potential for better care – if your doctor can access your full medical history, he or she can likely improve his or her effectiveness.

The Personal Data Ecosystem Consortium (PDEC) was also created in 2010 with the stated purpose of connecting start-up business to share personal data for reciprocal benefit among members. The PDEC also advocates for individuals to be empowered to access and utilize their own data. Such cooperatives are also appearing on a much larger scale. Industry lines are disappearing and corporations are teaming up, sharing data, and mutually improving their business models with more data analysis.

In the retail industry, billions of records are generated daily. Many retailers have long been using sales logistics as a means of data analysis. The data generated in sales was formerly reserved to advance the interests of only that retailer to whom the data belongs. In the modern commercial world, retailers exchange data to simultaneously track competitors' sales, which allows for a higher degree of data analysis.

An example of retail data monetization can be found in Target, a retailer in the United States with websites that can be used by domestic and international consumers. This retailer built a marketing strategy that used predictive data analysis as a means of trying to ascertain whether a customer was pregnant. By using consumer-created Target baby shower registry records, they were able to track purchases that many pregnant women had in common and use that to predict other customers who might be pregnant along with their anticipated due dates. This allowed the corporation to send coupons to non-registered expectant mothers. Target reasons that if they can secure new parents as customers in the second or third trimester of pregnancy, they can count on them as being loyal to Target's brand for at least a few years. In addition, the simple use of a credit card authenticates a Guest ID at Target. Target can then gather information on you through their records, or they can purchase data from other data creators. This can enable a retailer to use your purchasing history or real-time purchasing behavior to guide you to other products you might like or to entice you to make repeat purchases through targeted promotions or coupons. This is highly beneficial to corporations given the pace of commerce in the twenty-first century.

The financial services industry is another prime example of data monetization in action. Credit card companies have long been in the business of using transaction data for profit maximization and also selling such data.

This data is also a liability. Increasingly, financial institutions have become targets of cyber attacks, and cyber security has become an

increasing concern for businesses of all types. Cyber attacks can potentially steal credit card data for a multitude of customers, which can undercut consumer confidence or reduce future business from wary customers once the breach is made known to the consumer and broader public. In August 2014, the previously referenced Federal Bureau of Investigation (FBI) opened an investigation into hacking attacks on seven of the top 15 banks, most notably JPMorgan Chase. The FBI is still unsure as to the nature or origin of the attack, and it is also unclear whether the hackers accessed consumer banking or investment accounts. Around Thanksgiving the year prior, megaretailer Target was hit with a similarly unprecedented cyber attack. In this instance, hackers may have accessed information from as many as 110 million customers. The origin of that cyber attack remains unknown. What is known of the attack is that over 40 million credit card numbers were stolen.

Further Reading

- About Us. *Personal Data Ecosystem Consortium*. <http://pde.cc/aboutus/>.
- Data Mining and Profiling. SAGE encyclopedia on Big Data.
- Data Monetization in the Age of Big Data. *Accenture*. <http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Data-Monetization-in-the-Age-of-Big-Data.pdf>.
- FBI Expands Ability to Collect Cellphone Location Data, Monitor Social Media, Recent Contracts Show. *The Intercept*. <https://theintercept.com/2020/06/24/fbi-surveillance-social-media-cellphone-dataminr-venntel/>.
- FBI investigating hacking attack on JPMorgan. *CNN Money*. <http://money.cnn.com/2014/08/27/investing/jpmorgan-hack-russia-putin/>.
- How Companies Learn Your Secrets. *The New York Times*. <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&r=0>.
- Target: Hacking hit up to 110 Million Customers. *CNN Money*. <http://money.cnn.com/2014/01/10/news/companies/target-hacking/>.

Data Monitoring

► Data Profiling

Data Munging and Wrangling

Scott N. Romaniuk

University of South Wales, Pontypridd, UK

The terms “Data Munging” and “Data Wrangling” (also refers to “data cleaning”) are common terms in the world of programmers and researchers. They are interchangeable and refer to the manual conversion of raw data into another form that makes it easier for the programmer, researcher, and others to understand and to work with. This process also involves what is referred to as the “mapping” of raw forms of data or data files (e.g., txt, csv, xml, and json), and applying it to another format.

During the course of performing data analysis and visualization, the performance of which are referred to as “data science,” researchers often face the creation of messy data sets, and this is especially the case with larger and more complex data sets. Data munging, therefore, describes the process of sorting through either small or large data sets, which can become messy and disorderly, and “cleaning” it up or manipulating it. This process is often accomplished with the aim of creating a final or conclusive form of data, and data presentation or recognition. After cleaning the data, it can then be used more efficiently and lend itself to other uses. It can also involve the manipulation of multiple data sets. Simplified, the primary steps involved in data munging are:

- Addressing variable and observation names (rows and columns) including the creation of new variables that might be required
- Consolidate data into a single unified data set
- Molding or shaping data (address missing data/values, dropping data, dealing with outliers, balance the data/ensure consistency)

The creation of “tidy data” is important for handling data and moving them between programs and sharing them with others and can often be a painstaking process, but one that is

critical for working efficiently and effectively with data to be analyzed. The idea of “tidy data” refers to the ease with which data and data sets can be navigated. Some of the core features include the placement of data such as variables in rows and columns, the removal of errors in the data, ensuring internal consistency, and ensuring that data has been converted into complementary formats. Language harmonization falls under the category of “tidy data,” including the synchronization of communication elements so that variables of the same “type” can be grouped together. For example, “men” and “males,” and “women” and “females,” representing the same two populations can be grouped using a single label for both. This can be applied to a range of items that share similar enough features or conditions that they can be grouped accordingly.

Data munging can be performed through the use of a variety of software and tools, including but not limited to Pandas, R, Stata, and SPSS, all of which present the programmer or researcher with useful data manipulation tools or capabilities. Python is one of the most popular Python packages for creating and managing data structures. While data sets can be “messy,” the term does not necessarily imply that a given data set is not useful or has not been created properly. A “messy” data set might be difficult to work with and therefore requires some preprocessing technique in preparation for further work or presentation, depending on what purpose is intended for the data.

A programmer or researcher, for example, can: (a) correspond different data values so they are formatted the same way; (b) delete or harmonize data vocabulary so that data are easier to locate, and address the issue of missing values in data sets; (c) transfer data sets from one programming tool or package to another, removing certain values that are either irrelevant to the research or information that the programmer or researcher wants to present, and separate or merge rows and columns.

Compiling data often results in the creation of data sets with missing values. Missing values can lead to problems of representation or internal and external validity in research, which is an

important issue to address in work using regression. Missing data results in difficulty in determining the impact on regression coefficients. Proceeding with data analysis using data sets with missing values can also lead to a distortion within the analysis, or bias, and an overall less-desirable quality of work.

For example, during the course of field research involving the distribution of questionnaires, more than half of the questionnaires returned with unchecked boxes can result in the misrepresentation of a given condition being studied. Values can be missing for various reasons, including responses improperly recorded and even a lack of interest on the part of the individual filling out the questionnaire. It is, therefore, necessary to look at the type of missing values that exist in a given data set.

In order to address this problem, the programmer or researcher will undertake what is called a “filtering” process. “Filtering” in this context refers simply to the deliberate removal of data, resulting in a tidier dataset. The process can also involve sorting data to present or highlight different aspects of the dataset. For example, if a researcher examines a set number of organizations in a particular city or country, they may want to organize the responses by a specific organization type or category. Categorization in this sense can be made along the lines of gender, type of organization, or area(s) of operation. Organizing the data in this way can also be referred to as an isolation process.

This process is useful for exploratory and descriptive data analysis and investigation. If a researcher is interested in examining country support for political issues, such as democracy, countries can be categorized along, for example, region or regime type. In doing so, the programmer or researcher can illustrate which regime type (s) is/are more likely to be interested in politics or political issues. Another example can involve the acquisition of data on the number of people who smoke or consume alcohol excessively in a society. Data could then be sorted according to gender, age, or location, and so on.

Data munging may involve the categorical arrangement or grouping of information. It is possible to determine if there is a relationship

between two or more variables or conditions. Using the previous examples, a researcher may want to determine if there exists any correlation between gender and smoking, or country type and interest in political issues or involvement in political processes. Various groupings can also be performed to determine the existence of further correlations. Doing so can lead to the creation of a more robust data structure.

Data munging as a traditional method, has been referred to as an “old(er)” and “outdated” process, given that it was invented decades ago. Recently, more streamlined and integrated methods of data cleaning or arrangement have been formulated and are now available, such as Power Query. Data munging can take a great deal of time and since the process is a manual one, the outcome of the process can still contain errors. A single mistake at any stage of the process can result in subsequent and unintended errors produced due to the initial error. Nonetheless, data munging constitutes an important part of the data handling and analysis process and is one that requires careful attention to detail as is the case with other types of research. The practice of treating data in ways discussed is akin to filling holes in a wall, smoothing the surface and applying primer before painting. It is a critical step in preparing data for further use and will ultimately aid the researcher working with data in various quantities.

Cross-References

- [Data Aggregation](#)
- [Data Cleansing](#)
- [Data Integrity](#)
- [Data Processing](#)
- [Data Quality Management](#)

Further Reading

- Clark, D. (2020). Data munging with power query. In *Beginning Microsoft Power BI*. Berkeley: Apress. https://doi.org/10.1007/978-1-4842-5620-6_3.
- Endel, F., & Piringer, H. (2015). Data wrangling: Making data useful again. *IFAC-PapersOnLine*, 48(1),

111–112. <https://www.sciencedirect.com/science/article/pii/S2405896315001986>.

Foxwell, H. J. (2020). Cleaning your data. In *Creating good data*. Berkeley: Apress. https://doi.org/10.1007/978-1-4842-6103-3_8.

<https://www.elderresearch.com/blog/what-is-data-wrangling-and-why-does-it-take-so-long/>.

Skiena, S. S. (2017). Data munging. In *The data science design manual. Texts in computer science*. Cham: Springer. https://doi.org/10.1007/978-3-319-55444-0_3.

Thurber, M. (2018, April 6). What is data wrangling and why does it take so long? Elder Research. <https://www.elderresearch.com/blog/what-is-datawrangling-and-why-does-it-take-so-long/>.

Wähner, K. (2017, March 5). Data preprocessing vs. Data wrangling in machine learning projects. InfoQ. infoq.com/articles/ml-data-processing/.

Wiley, M., & Wiley, J. F. (2016). Data munging with Data. Table. In *Advanced R*. Berkeley: Apress. https://doi.org/10.1007/978-1-4842-2077-1_8.

Data Pre-processing

- [Data Cleansing](#)

Data Preservation

- [Data Provenance](#)

Data Privacy

- [Anonymization Techniques](#)

Data Processing

Fang Huang
Tetherless World Constellation, Rensselaer
Polytechnic Institute, Troy, NY, USA

Synonyms

[Data discovery](#); [DP](#); [Information discovery](#); [Information extraction](#)

Introduction

Data processing (DP) refers to the extraction of information through organizing, indexing, and manipulating data. Information here means valuable relationships and patterns that can help solve problems of interest. In *history*, the capability and efficiency of DP have been improving with the advancement of technology. Processing involving intensive human labor has been gradually replaced by machines and computers. The methods of DP refer to the techniques and algorithms used for extracting information from data. For example, processing of facial recognition data needs classification, and processing of climate data requires time series analysis. The results of DP, i.e., the information extracted, also depend largely on *data quality*. Data quality problems like missing values and duplications can be solved through various methods, but some systematic issues like equipment design error and data collection bias are harder to overcome at this stage. All these aspects influencing DP will be covered in later sections, but let's look at the history of DP first.

History

With the advancement of technology, the history of DP can be divided into three stages: Manual DP, Mechanical DP, and Electronic DP. The goal is to finally use Electronic DP to replace the other two to reduce error and improve efficiency.

- *Manual DP* is to process data with little or no aid from machines. Before the stage of Mechanical DP, only small-scale DP could be done, and they were very slow and could easily bring in errors. Having said that, Manual DP still exists at present, and it is usually because the data are hard to digitize or not machine readable, like in the case of retrieving sophisticated information from old books or records.
- *Mechanical DP* is to process data with help from mechanical devices (not modern computers). This stage started in 1890 (Bohme et al. 1991). During that year, the US Census Bureau installed a system which consists of complicated punch card machines to help tabulate the results of a recent national census of population. All data are organized, indexed, and classified. Searching and computing are made easier and faster than manual work with the punch card system.
- *Electronic DP* is to process data using computer and other advanced electronic devices. Nowadays, Electronic DP is the most common method and is still quickly evolving. It is widely seen in online banking, ecommerce, scientific computing, and other activities. Electronic DP provides best accuracy and speed. Without specification, all DP discussed in later sections are Electronic DP.

Methods

The methods of DP here refer to the techniques and algorithms used for extracting information from data, which vary a lot with the information of interest and data types. One definition for data is “Data are encodings that represent the qualitative or quantitative attributes of a variable or set of variables” (Fox 2018). The categorical type of data can represent qualitative attributes, like the eye color, gender, and jobs of a person; and the quantitative attributes can be represented by numerical type of data, like the weight, height, and salary of a person. Based on data types and patterns of interest, we can choose from several DP methods (Fox 2018), such as:

- *Classification* is a method that uses a classifier to put unclassified data into existing categories. The classifier is trained using categorized data labeled by experts, so it is one type of supervised learning in the machine learning terminology. Classification works well with categorical data.
- *Regression* is a method to study the relationship between a dependent variable and other independent variables. The relationship can be

used for predicting future results. Regression usually uses numerical data.

- *Clustering* is a method to find distinct groups of data based on their characteristics. Social media companies usually use it to identify people with similar interests. It is a type of unsupervised learning and works with both qualitative and quantitative data.
- *Association Rule Mining* is a method to find relationships between variables such as “which things or events occur together frequently in the dataset?”. This method was initially developed to do market basket analysis. Researchers in mineralogy apply association rule mining to use one mineral as the indicator of other minerals.
- *Outlier Analysis*, also called anomaly detection, is a method to find data items that are different from the majority of the data.
- *Time Series Analysis* is a set of methods to detect trends and patterns from the time series data. Time series data are a set of data indexed with time, and this type of data are used in many different domains.

DP methods are a key part for generating information, and the abovementioned ones are just a partial list. Data quality is another thing that can influence the results of DP.

Data Quality

Thota (2018) defined data quality as “the extent to which some data successfully serve the purposes of the user.” In the case of DP, this purpose is to get correct information, which involves two levels of data quality. Level 1 is the issues in data themselves such as missing values, duplications, inconsistency among data, and so on. Level 2 is accuracy, which is the distance between data and real values. Usually level 1 quality problems are easier to solve compared to the level 2.

Level 1 issues can be discovered through exploratory inspection and visualization tools, and the issues can be solved accordingly.

- *Missing data*: direct solution is to go back and gather more information. If that is not possible, common solution is to use domain knowledge or existing algorithms to impute them based on correlations between different attributes.
- *Duplications*: we can index the data based on certain attributes to find out duplications and remove them accordingly.
- *Inconsistency*: check the metadata to see the reason for inconsistency. Metadata are supplementary descriptive information about the data. Common issues are inconsistent data units and formats, which can be converted accordingly.
- *Outliers*: if they are proven to be errors, we can make corrections. In other cases, outlier data should still be included.

Level 2 issues are usually originated from equipment design error and/or data collection bias, in which case data are uniformly off the true values. This makes level 2 issues harder to be seen in data validation. They can possibly be seen when domain experts check the results.

Data Processing for Big Data

Big data are big in three aspects: volume, velocity, and variety (Hurwitz et al., 2013). Variety has been discussed in data quality section. Now we need to have a model capable of storing and processing high volume of data at relatively fast speed and/or to deal with fast and continuous high-speed incoming data in a short response time.

Batch processing works with extremely large static data. In this case, there is no new data coming in during processing, and data usually stores in more than one device. The large dataset is grouped into smaller batches and processed, respectively, with results combined later. The processing time is relatively long, so this model is not suitable for real-time tasks.

Stream processing, in contrast to batch processing, applies to dynamic and continuous data (new data keeps coming in). This model usually works for tasks that require short response time. For example, hotel and airline reservation systems need to provide instant feedbacks to customers. Theoretically, this model can handle unlimited amount of data as long as the system has enough capacity.

Mixed processing could process both batch and stream data.

Big Data Processing Frameworks

A framework in computer science is a special library of generic programs, which can perform specific tasks after adding some codes for actually functionality. Widely used frameworks are usually well written and tested. Simple python scripts can take care of small datasets, but for big data systems, "...processing frameworks and processing engines are responsible for computing over data in a data system" (Ellingwood 2016). This section will cover five popular open-source big data processing frameworks from Apache.

1. *Apache Hadoop*: As a well-tested batch-processing framework, Apache Hadoop was first developed by Yahoo to build a "search engine" to compete with Google. Later Yahoo found its great potential in DP. Hadoop's cross-system compatibility, distributed system architecture, and open-source policy made it popular with developers. Hadoop can easily scale up from one individual machine (for testing and debugging) up to data servers with large number of nodes (for large-scale computing). Due to its distributed architecture, a high-workload system can be extended by adding new nodes, and batch process can be made more efficient through parallel computing (Ellingwood 2016). Major components of Hadoop are:

Hadoop Distributed File System (HDFS):

HDFS is the underlying file managing and coordinating system that ensures efficient data file communication across all nodes.

Yet Another Resource Negotiator (YARN): YARN is the manager to schedule tasks and manage system resources.

MapReduce: MapReduce is the programming model taking advantage of "divide and conquer" algorithm to speed up DP.

2. *Apache Storm*: Apache Storm is a stream processing framework suitable for highly responsive DP. In a stream processing scenario, data coming in the system are continuous and unbounded. To achieve the goal of delivering results in nearly real time, Storm will divide the incoming data stream into small and discrete units (smaller than batch) for processing. These discrete steps are called bolts. Native Storm does not keep operations on bolts in order, which has to be solved by adding extra modules like Trident (Ellingwood 2016). The bottom line is, Storm is highly efficient and supports multiprogramming languages, so it is suitable for low-latency stream processing tasks.
3. *Apache Samza*: "Apache Samza is a stream processing framework that is tightly tied to the Apache Kafka messaging system." (Ellingwood 2016). Apache Kafka is a distributed streaming platform that process streams in the order of occurrence time and keep an immutable log. Hence, Samza can natively overcome the ordering problem in Storm and enable real-time collaboration on DP between multiple teams and applications in big organizations. However, compared to Storm, Samza has higher latency and less flexibility in programming language (only support Java and Scala).
4. *Apache Spark*: Apache Spark is the next-generation framework that combines batch and stream processing capabilities. Compared to Hadoop MapReduce, Spark processes data much faster due to its optimization on in-memory processing and task scheduling. Furthermore, the deployment of Spark is more flexible – it can run individually on a single system or replace the MapReduce engine and incorporate into a Hadoop system. Beyond

that, Spark programs are easier to write because of an ecosystem of existing libraries. It generally does a better job in batch processing than Hadoop; however, it might not fit for extremely low-latency stream processing tasks. Also, devices using Spark need to install larger RAM, which increases costs. Spark is a versatile framework that fits diverse processing workloads.

5. *Apache Flink*: Apache Flink framework handles both stream and batch processing workloads. It simply treats batch tasks as bounded stream data. This stream-only approach offers Flink fast processing speed and real in order processing. It is probably the best choice for organizations with strong needs for stream processing and some needs for batch processing. Flink is relatively flexible because it is compatible with both Storm and Hadoop. However, its scaling capability is somewhat limited due to its short history.

Many open-source big data processing systems are available on the market, and each has its own strengths and drawbacks. There is no “best” framework and no single framework that can address all user needs. We need to choose the right one or combine multiple frameworks based on the needs of real-world projects.

Further Reading

- Bohme, F., Wyatt, J. P., & Curry, J. P. (1991). *100 years of data processing: the punchcard century (Vol. 3)*. US Department of Commerce, Bureau of the Census, Data User Services Division.
- Ellingwood J. (2016). Hadoop, storm, samza, spark, and flink: Big data frameworks compared. Retrieved 25 Feb 2019, from <https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared>.
- Fox, P. (2018). Data analytics course. Retrieved 25 Feb 2019, from <https://tw.rpi.edu/web/courses/DataAnalytics/2018>.
- Hurwitz, J. S., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big data for dummies*. Hoboken: Wiley.
- Thota, S. (2018). *Big data quality*. Springer: Encyclopedia of Big Data.

Data Profiling

Patrick Juola

Department of Mathematics and Computer Science, McNulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Synonyms

Data monitoring

Data profiling is the systematic analysis of a data source, typically prior to any specific use, to determine how useful it is and how best to work with it. This analysis will typically address matters such as what information the source contains, the meta-data about the information, the quality of data, whether or not there are issues such as missing or erroneous elements, and patterns present within the data that may influence its use. Data profiling helps identify and improve data quality, which in turn improves the information systems built upon them (Azeroual et al. 2018).

Errors are inevitable in any large human project. Particularly in collecting big data, some typical types of errors include (Azeroual et al. 2018):

- Missing data
- Incorrect information
- Duplicate data
- Inconsistently represented data

For example, the “telephone number” column of a large customer database should be expected to contain telephone numbers (Kimball 2004). In the United States and Canada, such a number is defined by ten numeric digits but might be stored as a string. Empty cells may or may not represent genuine data (where the customer refused to provide a number) but may also represent data entry errors. Entries with alphanumeric characters may be marketing schemes but are likely to be errors. Even correct data may be duplicated or represented inconsistently (212-555-1234, (212) 555-1234, and +1 212 555 1234 are the same

number, written differently; similarly, Scranton, PA, Scranton, Penna., and Scranton, Pennsylvania are the same city).

In addition to actual errors, data can be profiled for bias and representativeness. Depending upon how the data is collected, not all elements of the relevant universe will be equally sampled, which in turn will bias inferences drawn from the database and reduce their accuracy.

The data profiling process will typically involve reviewing both the structure and content of the data. It will confirm that the data actually describe what the metadata say they should (e.g., that the “state” column contains valid states and not telephone numbers). It should further identify relationships between columns and should label problematic outliers or anomalies in the data. It should confirm that any required dependencies hold (for example, the “date-of-birth” should not be later than the “date-of-death”). Ideally, it may be possible to fix some of the issues identified or to enhance the data by using additional information. If nothing else, data profiling will provide a basis for a simple “Go–No Go” decision about whether the proposed project can go forward or about whether the proposed database is useful (Kimball 2004). If the needed data are not in the database, or the data quality is too low, it is better to learn (early) that the project cannot continue.

In addition to improving data quality, data profiling also can be used for data exploration, by providing easily understandable summaries of new datasets. Abedjan et al. (2015) provide as examples “files downloaded from the Web, old database dumps, or newly gained access to some [databases]” with “no known schema, no or old documentation, etc.” Learning what is actually stored in these databases can enhance their usefulness. Similarly, data profiling can be used to help optimize a database or even to reverse-engineer it. Perhaps most importantly, data profiling can ease the task of integrating several independent databases to help solve a larger problem to which they are all relevant.

Data profiling is, therefore, as argued by Kimball (2004), a highly valuable step in any data-warehousing project.

Cross-References

- [Data Quality Management](#)
- [Metadata](#)

Further Reading

- Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: A survey. *The VLDB Journal*, 24, 557–581. <https://doi.org/10.1007/s00778-015-0389-y>.
- Azeroual, O., Saake, G., & Schallehn, E. (2018). Analyzing data quality issues in research information systems via data profiling. *International Journal of Information Management*, 41, 50–56., ISSN 0268-4012. <https://doi.org/10.1016/j.ijinfomgt.2018.02.007>.
- Kimball, R. (2004). *Kimball design tip #59: Surprising value of data profiling*. Number 59, September 14, 2004. <http://www.kimballgroup.com/wpcontent/uploads/2012/05/DT59SurprisingValue.pdf>.

Data Provenance

Ashiq Imran¹ and Rajeev Agrawal²

¹Department of Computer Science & Engineering, University of Texas at Arlington, Arlington, TX, USA

²Information Technology Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, USA

Synonyms

[Big data](#); [Cybersecurity](#); [Data preservation](#); [Data provenance](#); [Data security](#); [Data security management](#); [Metadata](#); [Privacy](#)

Introduction

Data provenance refers to the description of the origin, creation, and propagation process of data. Provenance is the lineage and the derivation of the data, documented history of an object, in other words, how the object was created, modified, propagated, and disseminated to its current location/status. By observing the provenance of an

object, we can infer the trustworthiness of the object. It stores ownership and process history about data objects. Provenance has been studied extensively in the past and people usually use provenance to validate physical objects in arts, literary works, manuscript, etc. Recently, the domain for provenance has gained significant attention in digital world and e-science. The provenance of data is crucial for validating, debugging, auditing, evaluating the quality of data and determining reliability in data. In today's period of Internet, complex ecosystems of data are even more ubiquitous. Provenance has been typically considered by the database, workflows, and distributed system communities. Capturing provenance can be a burdensome and labor-intensive task.

With the growing inundation of scientific data, a detailed description of metadata is important to share data and find the data and scientific results for the scientists. Scientific workflows assist scientists and programmers with tracking their data through all transformations, analyses, and interpretations. Data sets become trustworthy when the process used to create them are reproducible and analyzable for defects. Current initiatives to effectively manage, share, and reuse ecological data are indicative of the increasing importance of data provenance. Examples of these initiatives are National Science Foundation Datanet projects, Data Conservancy, DataONE.

Big data concept refers to a database which is continuously expanding and slowly becomes difficult to control and manage. The difficulty can be related to data capture, storage, search, sharing, analytics, and visualization, etc. Provenance in big data has been identified by a recent community whitepaper on the challenges and opportunities of big data.

Provenance has found applications in debugging data, trust, probabilistic data, and security (Hasan et al. 2007; Agrawal et al. 2014). Data provenance may be critical for applications with typical big data features (volume, velocity, variety, value, and veracity). A usual approach to handle the velocity aspect of big data is to apply data cleaning and

integration steps in a pay-as-you-go fashion. This has the advantage of increasing the timeliness of data, but in comparison with the traditional approach of data warehousing comes at the cost of less precise and less well-documented metadata and data transformations. Without information of provenance, it is difficult for a user to understand the relevance of data, to estimate or judge its quality, and to investigate unexpected or erroneous results. Big data systems that automatically and transparently keep track of provenance information would introduce pay-as-you-go analytics that do not suffer from this loss of important metadata. Moreover, provenance can be used to define meaningful access control policies for heavily processed and heterogeneous data. For instance, a user can be granted access to analysis results if they are based on data that person owns.

Big Data

Big Data is a buzzword used to describe the rapid growth of both structured and unstructured data. With the rapid development of social networking, data collection capacity, and data storage, big data are growing swiftly in all science and engineering domains including social, biological, biomedical sciences (Wang et al. 2015; Glavic 2014). Examples are Facebook Data, Twitter Data, Linked-In Data, and Health Care Data.

In simple words, big data can be defined as data that is too big, too quick, or too hard for existing tools to process and analyze. Here, “too big” means that organizations increasingly must deal with terabyte-scale or petabyte-scale collections of data. For example, *Facebook* generates and stores four images of different sizes, which translates to a total of 60 billion images and 1.5 PB of storage. “Too quick” means data is not only huge enough, but also it must be processed and analyzed quickly – for example, to identify fraud at a point of sale or transaction. Lastly, “too hard” means data may not follow any particular structure. As a result, no existing tool can process and analyze it properly. For example, data that is created in media, such as

MP3 audio files, JPEG images, and Flash video files, etc.

According to Weatherhead University Professor Gary King, “There is a big data revolution.” But the revolution is not about the quantity of data rather using the data and doing a lot of things with the data. To understand the phenomenon that is big data, it is often described using five Vs: Volume, Velocity, Variety, Veracity, and Value.

Provenance in Big Data

Big data provenance is a type of provenance to serve scientific computation and workflows that process big data. Recently, an interesting example has come up. Who is the most popular footballer in the world? From social media data, all the fans around the world select their favorite footballer. This generates a huge volume of data. This data carries the vote of the fans. Such a massive amount of data must be able to provide desired result.

Let’s consider a scenario. Whenever we find some data, do we think what the source of data is? It is quite possible that data is copied from somewhere else. It is also possible that data is incorrect. The data we usually see on the web such as rating of a movie or smart phone, news story. Do we think about it, how much legitimate is it? For scientists, they need to have confidence on accuracy and timeliness on the data that they are using.

Some of the common challenges of big data are as follows:

1. It is too difficult to access all the data.
2. It is difficult to analyze the data.
3. It is difficult to share information and insights with others.
4. Queries and reports take a long time to run.
5. Expertise needed to run the analysis legitimately.

Without provenance it is nearly impossible for a user to know the relevance of data, assess the quality of its data, and to explore the unexpected or erroneous result.

Application of Provenance

Provenance systems may be created to support a number of uses and according to Goble, various applications of provenance are as follows:

- **Data Quality:** Lineage can be used to estimate data quality and data reliability based on the source data and transformations. It can also provide proof statements on data derivation.
- **Audit Trail:** Provenance can be used to trace the audit trail of data and evaluate resource usage and identify errors in data generation.
- **Replication Recipes:** Thorough provenance information can allow repetition of data derivation, help maintain its currency, and be a recipe for replication.
- **Informational:** A generic use of lineage is to query based on lineage metadata for data discovery. It can also be browsed to provide a context to interpret data.

Provenance in Security

The fundamental parts of the security are the confidentiality, integrity, and availability. Confidentiality indicates protection of data against disclosure. Sensitive information such as commercial or personal information is necessary to keep confidential. Provenance information covers access control mechanism. With the progress of advanced software applications more complex security mechanisms must be used. Traditional access control mechanisms are built for specific purposes and are not easily configured to address the complex demands. If we are able to trace the dependency of access, then it will provide essential information of security.

Challenges

Information recording about the data at origin does not come into play unless this information can be interpreted and carried through data analysis. But there are a lot of issues of data provenance such as query inversion, uncertainty

of sources, data citation, and archives management. To acquire provenance in big data is a challenging task. Some of the challenges (Wang et al. 2015; Glavic 2014; Agrawal et al. 2014) are:

- **Uncommon Structure:** It is hard to define a common structure to model the provenance of data sets. Data sets can be structured or unstructured. Traditional databases and workflows may follow structured way but for big data it may not necessarily be true. We cannot reference separate entries in the file for provenance without knowing the way how the data is organized in a file.
- **Track Data of Distribute Storage:** Big data systems often distribute in different storages to keep track of data. This may not be applicable for traditional databases. For provenance, we need to trace and record data and process location.
- **Check Authenticity:** Data provenance needs to check in timely manner to verify authenticity of data. As increasing varieties and velocities of data, data flows can be irregular. Periodic or event triggered data loads can be challenging to manage.
- **Variety of Data:** There is variety of data such as unstructured text documents, email, video, audio, stock ticker data, and financial transactions which comes from multiple sources. It is necessary to connect and correlate relationships, hierarchies, and multiple data linkages otherwise data can quickly get out of control.
- **Velocity of Data:** We may need to deal with streaming data that comes at unprecedented speed. We need to react quickly enough to manage such data.
- **Lack of Expertise:** Since big data is fairly a new technology, there are not enough experts who know how to deal with big data.
- **Secure Provenance:** There is a huge volume of data and information. It is important to maintain privacy and security of provenance information. But, it will be challenging to maintain privacy and integrity of provenance information with big data.

Opportunities

The big secret of big data is not about the size of the data; it is about relevancy of data which is rather small in comparison. Timely access to appropriate analytic insights will replace the need for data warehouses full of irrelevant data, most of which could not be managed or analyzed anyway. There are a lot of opportunities of provenance (Wang et al. 2015; Glavic 2014; Agrawal et al. 2014), which are listed below:

- **Less Overhead:** We need to process huge volume of data, so high performance is critical. It is necessary that provenance collection has minor impact on the application's performance.
- **Accessibility:** A suitable coordination between the data and computer systems is required to access different types of big data for provenance and distributed systems.
- **User Annotations Support:** It is important to capture user notes or metadata. This is applicable for database and workflows as well as for big data. Thus, we need an interface that allows users to add their notes about the experiment.
- **Scalability:** Scalability comes into play for big data. The volume of big data is growing exponentially. Provenance data are also rapidly increasing which is making it necessary to scale up provenance collection.
- **Various Data Models Support:** So far data models for provenance system were structured such as database. It is important to support for unstructured and semi-structured provenance data models from users and systems because big data may not follow a particular structure.
- **Provenance Benchmark:** If we can manage to set up a benchmark for provenance, then we can analyze performance blockages and to compute performance metrics. Provenance information can be used to support data-centric monitoring.
- **Flexibility:** A typical approach to deal with velocity of the data is to introduce data cleaning and integration in pay-per-use fashion. This may reduce the cost and consume less time.

Conclusion

Data provenance and reproducibility of computations play a vital role to achieve improvement of the quality of research. Some studies have shown in the past that it is really hard to reproduce computational experiments with certainty. Recently, the phenomenon of big data makes this even harder than before. Some challenges and opportunities of provenance in big data are discussed in this article.

Further Reading

- Agrawal, R., Imran, A., Seay, C., & Walker, J. (2014, October). A layer based architecture for provenance in big data. In *Big Data (Big Data), 2014 I.E. international conference on* (pp. 1–7), IEEE.
- Glavic, B. (2014). Big data provenance: Challenges and implications for benchmarking. In *Specifying big data benchmarks* (pp. 72–80). Berlin/Heidelberg: Springer.
- Hasan, R., Sion, R., & Winslett, M. (2007, October). Introducing secure provenance: Problems and challenges. In *Proceedings of the 2007 ACM workshop on storage security and survivability* (pp. 13–18), ACM.
- Wang, J., Crawl, D., Purawat, S., Nguyen, M., & Altintas, I. (2015, October). Big data provenance: Challenges, state of the art and opportunities. In *Big Data (Big Data), 2015 I.E. international conference on* (pp. 2509–2516), IEEE.

Data Quality Management

Erik W. Kuiler

George Mason University, Arlington, VA, USA

Introduction

With the increasing availability of Big Data and their attendant analytics, the importance of data quality management has increased. Poor data quality represents one of the greatest hurdles to effective data analytics, computational linguistics, machine learning, and artificial intelligence. If the data are inaccurate, incomprehensible, or unusable, it does not matter how sophisticated

our algorithms and paradigms are, or how intelligent our “machines.”

J. M. Juran provides a definition of data quality that is applicable to current Big Data environments: “Data are of high quality if they are fit for their intended use in operations, decision making, and planning” (Juran and Godfrey 1999, p. 34.9). In this context, quality means that Big Data are relevant to their intended uses and are of sufficient detail and quantity, with a high degree of accuracy and completeness, of known provenance, consistent with their metadata, and presented in appropriate ways.

Big Data provide complex contexts for determining data quality and establishing data quality management. The Internet of Things (IoT) has complicated Big Data quality management by expanding the dynamic dimensions of scale, diversity, and rapidity that collectively characterize Big Data. From intelligent traffic systems to smart healthcare, IoT has inundated organizations with ever-increasing quantities of structured and unstructured Big Data sets that may include social media, public and private data sets, sensor logs, web logs, digitized records, etc., produced by different vendors, applications, devices, micro-services, and automated processes.

Conceptual Framework

Big Data taken out of their contexts are meaningless. As social constructs, Big Data, like “little data,” can only be conceptualized in the context of market institutions, societal norms, juridical constraints, and technological capabilities. Big Data-based assertions do not have greater claims to truth (however, one chooses to define this), objectivity, or accuracy than “small data-based” assertions.

Moreover, it should be remembered that just because Big Data are readily accessible does not mean that their uses are necessarily ethical. Big Data applications reflect the intersection of different vectors: technology, the maximizing the use of computational power and algorithmic sophistication and complexity; and analytics, the

exploration of very large data to formulate hypotheses and social, economic, and moral assertions.

Poor data quality presents a major hurdle to data analytics. Consequently, data quality management has taken on an increasingly important role within the overall framework of Big Data governance. It is not uncommon for a data analyst working with Big Data sets to spend approximately half of his or her time cleansing and normalizing data. Data of acceptable quality are critical to the effective operations of an organization and the reliability of its analytics and business intelligence. The application of Big Data quality dimensions and their attendant metrics, standards, as well as the use of knowledge domain-specific lexica and ontologies facilitate the tasks of Big Data quality management.

Dimensions of Big Data Quality

Big Data quality reflects the application of specific dimensions and their attendant metrics to data items to assess their acceptance for use. Commonly used dimensions of Big Data quality include:

- Accessibility – a data item is consistently available and replicable
- Accuracy – the degree to which a data item reflects a “real world” truth
- Completeness – the proportion of ingested and stored instance of a data item matches the expected input quantity
- Consistency – there are no differences between multiple representations of a data item and its stated definition
- Comparability – instances of a data item are consistent over time
- Correctness – a data item is error-free
- Privacy – a data item does not provide personally identifiable information (PII)
- Relevance – the degree to which a data item can meet current and future needs of its users
- Security – access to a data item is controlled to ensure that only authorized access can take place

Semiotic consistency – data items consistently use the same alphabet, signs, symbols, and orthographic conventions

Timeliness – a data item represents view of reality at a specific point in time

Trustworthiness – a data item has come from a trusted source and is managed reliably and securely

Uniqueness – a data item is uniquely identifiable so that it can be managed to ensure no duplication

Understandability – a data item is easily comprehended

Usability – a data item is useful to the extent that it may be readily understood and accessed

Validity – a data item is syntactically valid if it conforms to its stipulated syntax; a data item is semantically valid if it reflects its intended meaning

Data Quality Metrics

The function of measurements is to collect, calculate, and compare actual data with expected data.

Big Data quality metrics must exhibit three properties: they must be important to data users, they must be computationally sound, and they must be feasible. At a minimum, data quality management should reflect functional requirements and provide metrics of timeliness, syntactic conformance, semiotic consistency, and semantic congruence. Useful starting points for Big Data quality metrics include pattern recognition and identification of deviations and exceptions (useful for certain kinds of unstructured data) and statistics-based profiling (useful for structured data; for example, descriptive statistics, inferential statistics, univariate analyses of actual data compared to expected data).

Importance of Metadata Quality

Metadata describe the container as well as the contents of a data collection. Metadata support data interoperability and the transformation of Big Data sets into useable information resources. Because Big Data sets frequently come from

widely distributed sources, the completeness and quality of the metadata have direct effects on Big Data quality. Of particular importance to Big Data quality, metadata ensure that, in addition to delineating the identity, lineage and provenance of the data, the transmission and management of Big Data conform to predetermined standards, conventions, and practices that are encapsulated in the metadata and that access to, and manipulation of, the data items will comply with the privacy and security stipulations defined in the metadata. Operational metadata reflect the management requirements for data security and safeguarding personal identifying information (PII); data ingestion, federation, and integration; data anonymization; data distribution; and data storage. Bibliographical metadata provide information about the data item's producer, such as the author, title, table of contents, applicable keywords of a document. Data lineage metadata provide information about the chain of custody of a data item with respect to its provenance – the chronology of data ownership, stewardship, and transformations. Syntactic metadata provide information about data structures. Semantic metadata provide information about the cultural and knowledge domain-specific contexts of a data item.

Methodological Framework

Big Data quality management presents a number of complex, but not intractable, problems. A notional process for addressing these problems may include the following activities: data profiling, data cleansing, data integration, data augmentation, addressing issues of missing data.

Data Profiling

Data profiling provides the basis for addressing Big Data quality problems. Data profiling is the process of gaining an understanding of the data and the extent to which they comply with their quality specifications: are the data complete? are they accurate? There are many different

techniques and processes for data profiling; however, they can usually be classified in three categories:

Pattern analysis – expected patterns, pattern distribution, pattern frequency, and drill down analysis

Attribute analysis (e.g., data element/column metadata consistency) – cardinality, null values, ranges, minimum/maximum values, frequency distribution, and various statistics

Domain analysis – expected or accepted data values and ranges

Data Cleansing

Data profiling provides essential information for solving Big Data quality problems.

For example, the data profiling activities could reveal that the data set contains duplicate data items or that there are different representations of the same data item. This problem occurs frequently when merging data from different sources. Common approaches to solving such problems include:

Data exclusion – if the problem with the data is deemed to be severe, the best approach may be to remove the data

Data acceptance – if the error is within the tolerance limits for the data item, the best approach sometimes is to accept the data with the error

Data correction – if, for example, different variations of a data item occur, the best approach may be to select one version as the master and consolidate the different version with the master versions

Data value insertion – if a value for a field is not known and the data item is specified as NOT NULL, this problem may be addressed by creating a default value (e.g., unknown) and inserting that value in the field

Data Integration

Frequently in the integration of Big Data sets, it is not unusual to encounter problems that reflect the diverse provenance of the data items, in terms of

metadata specifications and cultural contexts. Thus, to ensure syntactic conformance and semantic congruence, the process of Big data integration may require parsing the metadata and data analyzing the contents of the data set accomplish the following:

- Metadata reconciliation – identity, categories, properties, syntactic and semantic conventions and norms
- Semiotic reconciliation – alphabet, signs, symbols, and orthographic conventions
- Version reconciliation – standardization of the multiple versions cross-referenced with an authoritative version

Data Augmentation

To enhance the utility of the Big Data items, it may be necessary augment a Big Data item with the corporation of additional external data to gain greater insight into the contents of the data set.

Missing Data

The topic of missing data has led to a lively discourse on imputation in predictive analytics. The goal is to conduct the most accurate analysis of the data to make efficient and valid inferences about a population or sample. Commonly used approaches to address missing data include:

- Listwise deletion – delete any case that has missing data for any bivariate or multivariate analysis
- Mean substitution – substitute the mean of the total sample of the variable for the missing values of that variable
- Hotdecking – identify a data item in the data set with complete data that is similar to the data item with missing data based on a correlated characteristic and use that value to replace the missing value in the other data item
- Conditional mean imputation (regression imputation) – use the equation to predict the values of the incomplete cases.
- Multiple-imputation analysis – based on the Bayesian paradigm; multiple imputation analysis has proven to be statistically valid from the frequentist (randomization-based) perspective.

K-nearest neighbor (KNN) – adapted from data mining paradigms: the mode of the nearest neighbor is used for discrete data; the mean is used for substituted for quantitative data

Challenges and Future Trends

The decreasing costs of data storage and the increasing rapidity of data creation and transportation assure the future growth of Big Data-based applications in both the private and public sectors. However, Big Data do not necessarily mean better information than that provided by little data. Big Data cannot overcome the obstacles presented by poorly conceived research designs or indifferently executed analytics. There remain Big Data quality issues that should be addressed. For example, many knowledge communities support multiple, frequently proprietary, standards, ontologies, and lexica, each of which with its own sect of devotees so that, rather than leading to uniform data quality, these proliferations tend have deleterious effects on Big Data quality and interoperability in global, cloud-based, IoT environments.

Further Reading

- Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, 67, 1012–1028.
- Allison, P. A. (2002). *Missing data*. Thousand Oaks: Sage Publications.
- Juran, J. M., & Godfrey, A. B. (1999). *Juran's quality handbook* (Fifth ed.). New York: McGraw-Hill.
- Labouseur, A. G., & Matheus, C. (2017). An introduction to dynamic data quality challenges. *ACM Journal of Data and Information Quality*, 8(2), 1–3.
- Little, R. J. A., & Rubin, D. B. (1997). *Statistical analysis with missing data*. New York: Wiley.
- Pipino, L. L. Y. W. L., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211–218.
- Saunders, J. A., Morrow-Howell, N., Spitznagel, E., Dore, P., Proctor, E. K., & Pescarino, R. (2006). Imputing missing data: A comparison of methods for social workers. *Social Work Research*, 30(1), 19–30.
- Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, 40(5), 103–110.
- Truong, H.-L., Murguzur, A., & Yang, E. (2018). Challenges in enabling quality of analytics in the cloud. *Journal of Data and Information Quality*, 9(2), 1–4.

Data Reduction

► Collaborative Filtering

Data Repository

Xiaogang Ma

Department of Computer Science, University of Idaho, Moscow, ID, USA

Synonyms

Data bank; Data center; Data service; Data store

Introduction

Data repositories store datasets and provide access to users. Content stored and served by data repositories includes digitized data legacy, born digital datasets, and data catalogues. Standards of meta-data schemas and identifiers enable the long-term preservation of research data as well as the machine accessible interfaces among various data repositories. Data management is a broad topic underpinned by the data repositories, and further efforts are needed to extend data management to research provenance documentation. The value of datasets is extended through connections from datasets to literature as well as inter-citations among datasets and literature.

Repository Types

A data repository is a place where datasets can be stored and accessed. Normally, datasets in a repository are marked up with metadata which provide essential context information about the datasets and enable efficient data search. The architecture of a data repository is comparable to a conventional library and a number of parallel comparisons can be found, such as datasets to publication hardcopies and metadata to

publication catalogues. While a conventional library needs certain physical space to store the hardcopies of publications, a data repository requests less physical resources, and it is able to organize various types of contents. In general, data repositories can be categorized into a few types based on the contents they organize. Among those types the mostly seen ones are digitized data legacy, born digital and data catalogue service, as well as the hybrid of them.

Massive amount of data have been recorded on media that are not machine readable, such as hardcopies of spreadsheets, books, journal papers, maps, photos, etc. Through the use of certain devices and technologies, such as image scanning and optical character recognition, those data can be transformed into machine-readable formats. The resulting datasets are part of the so-called digitized data legacy. Communities of certain scientific disciplines have organized activities for rescuing dataset from the literature legacy and improving their reusability and have developed data repositories to store the rescued datasets. For example, the EarthChem works on the preservation, discovery, access, and visualization of geoscience data, especially those in the fields of geochemistry, geochronology, and petrology (EarthChem 2016). A typical content type in EarthChem is spreadsheets which were originally published as tables in journal papers.

Comparing with digitized data legacy, more and more datasets are born digital since computers are increasingly used in data collection. A trend in academia is to use open-source formats for datasets stored in a repository and thus to improve the interoperability of the datasets. For example, comma-separated values (CSV) are recommended for spreadsheets. EarthChem is also open for born digital datasets. Recently, EarthChem has collaborated with publishers such as Elsevier to invite journal paper authors to upload the datasets used in their papers to EarthChem. Moreover, interlinks will be set up between the papers and the datasets. A unique type of born digital data is the crowdsourcing datasets. The contributors of crowdsourcing datasets are a large community of people rather than a few individuals or organizations. The

OpenStreetMap is such a crowdsourcing data repository for worldwide geospatial data.

The number of data repositories has significantly increased in recent years, as well as the subjects covered in those repositories. To benefit data search and discovery, another type of repositories has been developed to provide data catalogue service. For example, the data portal of the World Data System (WDS) (2016) allows retrieval of datasets from a wide coverage of WDS members through well-curated metadata catalogues encoded in common standards. Many organizations such as the Natural Environment Research Council in the United Kingdom and the National Aeronautics and Space Administration in the United States also provide data catalogue services to their data resources.

Data Publication and Citation

Many data repositories already mint unique identifiers such as digital object identifiers (DOIs) to registered datasets, which partially reflect people's efforts to make data as a kind of formal publication. The word data publication is derived from paper publication. If a journal paper is comparable to a registered dataset in a repository, then the repository is comparable to a journal. A paper has metadata such as authors, publication date, title, journal name, volume number, issue number, and page numbers. Most papers also have DOIs which resolve to the landing web pages of those papers on their publisher websites. By documenting metadata, minting DOIs to registered datasets in a repository, the datasets are made similar to published papers. The procedure of data publication is already technically established in many data repositories.

However, data publication is not just a technical issue. There are also social issues to be considered, because data is not conventional regarded as a "first-class" product of scientific research. Many datasets were previously published as supplemental materials of papers. Although data repositories make it possible to publish data as stand-alone products, the science community

still needs more time to give data an equal position as paper. A few publishers recently released so-called data journals to publish short description papers for datasets published in repositories, which can be regarded as a way to promote data publication. Funding organizations have also taken actions to promote data as formal products of scientific research. For example, the National Science Foundation in United States now allows funding applicants to list data and software programs as products in their bio-sketches.

A data repository has both data providers and data users. Accordingly, there are issues to be considered for both data publication and data citation. If a registered dataset is tagged with metadata such as contributor, date, title, source, publisher, etc. and is minted a DOI, then it is intuitively citable just like a published journal paper. To promote common and machine-readable metadata items among data repositories, a global initiative, the DataCite, has been working on standards of metadata schema and identifier for datasets since 2009. For example, DataCite suggests five mandatory metadata items for a registered dataset: identifier, creator, title, publisher, and publication year. It also suggests a list of additional metadata items such as subject, resource type, size, version, geographical location, etc. The methodology and technology developed by DataCite are increasingly endorsed by leading data repositories across the world, which make possible a common technological infrastructure for data citation.

Various communities have also taken efforts to promote best practices in data citation, especially the guidelines for data users. The FORCE11 published the Joint Declaration of Data Citation Principles in 2013 to promote good research practice of citing datasets. Earlier than that, in 2012, the Federation of Earth Science Information Partners (2012) published Data Citation Guidelines for Data Providers and Archives, which offers more practical details on how a published dataset should be cited. For example, it suggests seven required elements to be included in a data citation: authors, release date, title, version, archive and/or distributor, locator/identifier, and access date/time.

Data Management and Provenance

Works on data repositories underpin another broader topic, data management, which in general is about what one will do with the data generated during and after a research. The academia is now facing a cultural change on data management. Many funding agencies such as the National Science Foundation in United States now require researchers include a data management plan in funding proposals. From the perspective of researchers, good data management increases efficiency in their daily work. The data publication, reuse, and citation enabled by the infrastructure of data repositories increase the visibility of individual works. Good practices on data management and publication drive the culture of open and transparent science and can lead to new collaborations and unanticipated discoveries.

Though data repositories provide essential facilities for data management, developing a data management plan can still be time-consuming as it is conventionally not included in a research workflow. However, it is now regarded as a necessary step to ensure the research data to be safe and useful for both the present and future. In general, a data management plan includes elements such as project context, data types and formats, plans for short- and long-term management, data sharing and update plans, etc. A number of organizations provide online tools to help researchers draft such data management plans, such as the DMPTool developed by the California Digital Library, the tool developed by the Integrated Earth Data Applications project at Columbia University, and the DMPonline developed by the Digital Curation Centre in the United Kingdom.

Efforts on standards of metadata schema and persistent identifier for datasets in data repositories are enabling the preservation of data as research products. Recently, the academia takes a further step to extend the topic of data management to context management or, in a short word, the provenance. Provenance is about the origin of something. In scientific works documenting provenance includes linking a range of observations and model output, research activities, people and

organizations involved in the production of scientific findings with the supporting datasets, and methods used to generate them (Ma et al. 2014; Mayernik et al. 2013). Provenance involves works on categorization, annotation, identification, and linking among various entities, agents, and activities. To reduce duplicated efforts, a number of communities of practice have been undertaken, such as the CrossRef for publications, DataCite for datasets, ORCID for researchers, and IGSN for physical samples.

The Global Change Information System is such a data repository that is enabled with the functionality of provenance tracking. The system is led by the United States Global Change Research Program and records information about people, organizations, publications, datasets, research findings, instruments, platforms, methods, software programs, etc., as well as the interrelationships among them. If a user is interested in the origin of a scientific finding, then he or she can use the system to track all the supporting resources. In this way, the provenance information improves the reproducibility and credibility of scientific results.

Value-Added Service

The value of datasets is reflected in the information and knowledge extracted from them and their applications to tackle scientific, social, and business issues. Data repositories, data publication and citation standards, data management plans, and provenance information form a framework enabling the storage and preservation of data. To facilitate data reuse, more efforts are needed for data curation, such as data catalogue service, cross-disciplinary discovery, and innovative approaches for pattern extraction.

Thomson Reuters released the Data Citation Index recently, which indexes the world's leading data repositories and connects datasets to related refereed publications indexed in the Web of Science. Data Citation Index provides access to an array of data across subjects and regions, which enables users to understand data in a comprehensive context through linked content and summary

information. The linked information is beneficial because it enables users to gain insights which are lost when datasets or repositories are viewed in isolation. The quality and importance of a dataset are reflected in the number of citations it receives, which is recorded by the Data Citation Index. Such citations, on the other hand, enrich the connections among research outputs and can be used for further knowledge discovery.

In 2011, leading web search engines Google, Bing, Yahoo!, and Yandex started an initiative called Schema.org. Its aim is to create and support a common set of schemas for structured data markup on web pages. Schema.org adopts a hierarchy to organize schemas and vocabularies of terms, which are to be used as tags to mark up web pages. Search engine spiders and other parsers can recognize those tags and record the topic and content of a web page. This makes it easier for users to find the right web pages through a search engine. A few data repositories such as the National Snow and Ice Data Center in the United States already carried out studies to use Schema.org to tag web pages of registered datasets. If this mechanism is broadly adopted, a desirable result is a data search engine similar to the publication search engine Google Scholar.

Cross-References

- [Data Discovery](#)
- [Data Provenance](#)
- [Data Sharing](#)
- [Data Storage](#)
- [Metadata](#)

Further Reading

- EarthChem. (2016). About Earthchem. <http://www.earthchem.org/overview>. Accessed 29 Apr 2016.
- Federation of Earth Science Information Partners. (2012). Data citation guidelines for data providers and archives. <http://commons.esipfed.org/node/308>. Accessed 29 Apr 2016.
- Ma, X., Fox, P., Tilmes, C., Jacobs, K., & Waple, A. (2014). Capturing provenance of global change information. *Nature Climate Change*, 4(6), 409–413.

Mayernik, M. S., DiLauro, T., Duerr, R., Metsger, E., Thessen, A. E., & Choudhury, G. S. (2013). Data conservancy provenance, context, and lineage services: Key components for data preservation and curation. *Data Science Journal*, 12, 158–171.

World Data System. (2016). Trusted data services for global science. <https://www.icsu-wds.org/organization/intro-to-wds>. Accessed 29 Apr 2016.

Data Resellers

- [Data Brokers and Data Services](#)

Data Science

Lourdes S. Martinez
School of Communication, San Diego State
University, San Diego, CA, USA

Data science has been defined as the structured study of data for the purpose of producing knowledge. Going beyond simply using data, data science revolves around extracting actionable knowledge from said data. Despite this definition, confusion exists surrounding the conceptual boundaries of data science in large part due to its intersection with other concepts, including big data and data-driven decision making. Given that increasingly unprecedented amounts of data are generated and collected every day, the growing importance of the data science field is undeniable. As an emerging area of research, data science holds promise for optimizing performance of companies and organizations. The implications of advances in data science are relevant for fields and industries spanning an array of domains.

Defining Data Science

The basis of data science centers around established guiding principles and techniques that help organize the process of drawing out information and insights from data. Conceptually,

data science closely resembles data mining, or a process relying on technologies that implement these techniques in order to extract insights from data. According to Dhar, Jarke, and Laartz, data science seeks to move beyond simply explaining a phenomenon. Rather its main purpose is to answer questions that explore and uncover actionable knowledge that informs decision making or predicts outcomes of interest. As such, most of the challenges currently facing data science emanate from properties of big data and the size of its datasets, which are so massive they require the use of alternative technologies for data processing.

Given these characteristics, data science as a field is charged with navigating the abundance of data generated on a daily basis, while supporting machine and human efforts in using big data to answer the most pressing questions facing industry and society. These aims point toward the interdisciplinary nature of data science. According to Loukides, the field itself falls inside the area where computer programming and statistical analysis converge within the context of a particular area of expertise. However, data science differs from statistics in its holistic approach to gathering, amassing, and examining user data to generate data products. Although several areas across industry and society are beginning to explore the possibilities offered by data science, the idea of what constitutes data science remains nebulous.

Controversy in Defining the Field

According to Provost and Fawcett, one reason why data science is difficult to define relates to its conceptual overlap with big data and data-driven decision making. Data-driven decision making represents an approach characterized by the use of insights gleaned through data analysis for deciding on a course of action. This form of decision making may also incorporate varying amounts of intuition, but does not rely solely on it for moving forward. For example, a marketing manager faced with a decision about how much promotional effort should be invested in a

particular product has the option of solely relying on intuition and past experiences, or using a combination of intuition and knowledge gained from data analysis. The latter represents the basis for data-driven decision making. At times, however, in addition to enabling data-driven decision making, data science may also overlap with data-driven decision making. The case of automated online recommendations of products based on user ratings, preferences, and past consumer behavior is an example of where the distinction between data science and data-driven decision making is less clear.

Similarly, differentiating between the concepts of big data and data science becomes murky when considering that approaches used for processing big data overlay with the techniques and principles used to extract knowledge and espoused by data science. This conceptual intersection exists where big data technologies meet data mining techniques. For example, technologies such as Apache™ Hadoop® which are designed to store and process large-scale data can also be used to support a variety of data science efforts related to solving business problems, such as fraud detection, and social problems, such as unemployment reduction. As the technologies associated with big data are also often used to apply and bolster approaches to data mining, the boundary between where big data ends and data science begins continues to be imprecise.

Another source of confusion in defining data science stems from the absence of formalized academic programs in higher education. The lack of these programs exists in part due to challenges in launching novel programs that cross disciplines and the natural pace at which these programs are implemented within the academic environment. Although several institutions within higher education now recognize the importance of this emerging field and the need to develop programs that fulfill industry's need for practitioners of data science, the result up to now has been to leave the task for defining the field to data scientists.

Data scientists currently occupy an enviable position as among the most coveted employees for twenty-first-century hiring according to

Davenport and Patil. They describe data scientists as professionals, usually of senior-level status, who are driven by curiosity and guided by creativity and training to prepare and process big data. Their efforts are geared toward uncovering findings that solve problems in both private and public sectors. As businesses and organizations accumulate greater volumes of data at faster speeds, Davenport and Patil predict the need for data scientists will to continue in a very steep and upward trajectory.

Opportunities in Data Science

Several sectors stand to gain from the explosion in big data and acquisition of data scientists to analyze and extract insights from it. Chen, Chiang, and Storey note the opportunities inherent through data science for various areas. Beginning with e-commerce and the collection of market intelligence, Chen and colleagues focus on the development of product recommendation systems via e-commerce vendors such as Amazon that are comprised of consumer-generated data. These product recommendation systems allow for real-time access to consumer opinion and behavior data in record quantities. New data analytic techniques to harness consumer opinions and sentiments have accompanied these systems, which can help businesses become better able to adjust and adapt quickly to needs of consumers. Similarly, in the realm of e-government and politics, a multitude of data science opportunities exist for increasing the likelihood for achieving a range of desirable outcomes, including political campaign effectiveness, political participation among voters, and support for government transparency and accountability. Data science methods used to achieve these goals include opinion mining, social network analysis, and social media analytics.

Public safety and security represents another area that Chen and colleagues observe has prospects for implementing data science. Security remains an important issue for businesses and organizations in a post-September 11th 2001 era. Data science offers unique opportunities to provide additional protections in the form of security

informatics against terrorist threats to transportation and key pieces of infrastructure (including cyberspace). Security informatics uses a three-pronged approach coordinating organizational, technological, and policy-related efforts to develop data techniques designed to promote international and domestic security. The use of data science techniques such as crime data mining, criminal network analysis, and advanced multilingual social media analytics can be instrumental in preventing attacks as well as pinpointing whereabouts of suspected terrorists.

Another sector flourishing with the rise of data science is science and technology (S&T). Chen and colleagues note that several areas within S&T, such as astrophysics, oceanography, and genomics, regularly collect data through sensor systems and instruments. The result has been an abundance of data in need of analysis, and the recognition that information sharing and data analytics must be supported. In response, the National Science Foundation (NSF) now requires the submission of a data management plan with every funded project. Data-sharing initiatives such as the 2012 NSF Big Data program are examples of government endeavors to advance big data analytics for science and technology research. The iPlant Collaborative represents another NSF-funded initiative that relies on cyber infrastructure to instill skills related to computational techniques that address evolving complexities within the field of plant biology among emerging biologists.

The health field is also flush with opportunities for advances using data science. According to Chen and colleagues, opportunities for this field are rising in the form of massive amounts of health- and healthcare-related data. In addition to data collected from patients, data are also generated through advanced medical tools and instrumentation, as well as online communities formed around health-related topics and issues. Big data within the health field is primarily comprised of genomics-based data and payer-provider data. Genomics-based data encompasses genetic-related information such as DNA sequencing. Payer-provider data comprises information collected as part of encounters or exchanges between patients and the healthcare system, and includes

electronic health records and patient feedback. Despite these opportunities, Miller notes that application of data science techniques to health data remains behind that of other sectors, in part due to a lack of initiatives that leverage scalable analytical methods and computational platforms. In addition, research and ethical considerations surrounding privacy and protection of patients' rights in the use of big data present some challenges to full utilization of existing health data.

Challenges to Data Science

Despite the enthusiasm for data science and the potential application of its techniques for solving important real-world problems, there are some challenges to full implementation of tools from this emerging field. Finding individuals with the right training and combination of skills to become data scientists represents one challenge. Davenport and Patil discuss the shortage of data scientists as a case in which demand has grossly exceeded supply, resulting in intense competition among organizations to attract highly sought-after talent.

Concerns related to privacy represent another challenge to data science analysis of big data. Errors, mismanagement, or misuse of data (specifically data that by its nature is traceable to individuals) can lead to potential problems. One famous incident involved Target correctly predicting the pregnancy status of a teenaged girl before her father was aware of the situation, resulting in wide media coverage over issues equating big data with "Big Brother." This perception of big data may cause individuals to become reluctant to provide their information, or choose to alter their behavior when they suspect they are being tracked, potentially undermining the integrity of data collected.

Data science has been characterized as a field concerned with the study of data for the purpose of gleaning insight and knowledge. The primary goal of data science is to produce knowledge through the use of data. Although this definition provides clarity to the conceptualization of data science as a field, there persists confusion as to how data science differs from related concepts

such as big data and data-driven decision making. The future of data science appears very bright, and as the amount and speed with which data is collected continues to increase, so too will the need for data scientists to harness the power of big data. The opportunities for using data science to maximize corporate and organizational performance cut across several sectors and areas.

Cross-References

- [Big Data](#)
- [Big Data Research and Development Initiative \(Federal, U.S.\)](#)
- [Business Intelligence Analytics](#)
- [Data Mining](#)
- [Data Scientist](#)
- [Data Storage](#)
- [Data Streaming](#)

Further Reading

- Chen, H. (2006). *Intelligence and security informatics for international security: Information sharing and data mining*. New York: Springer Publishers.
- Chen, H. (2009). AI, E-government, and politics 2.0. *IEEE Intelligent Systems*, 24(5), 64–86.
- Chen, H. (2011). Smart health and wellbeing. *IEEE Intelligent Systems*, 26(5), 78–79.
- Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
- Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 90, 70–76.
- Dhar, V., Jarke, M., & Laartz, J. (2014). Big data. *Business & Information Systems Engineering*, 6(5), 257–259.
- Hill, K. (2012). How target figured out a teen girl was pregnant before her father did. *Forbes magazine*. *Forbes Magazine*.
- Loukides, M. (2011). *What is data science? The future belongs to the companies and people that turn data into products*. Sebastopol: O' Reilly Media.
- Miller, K. (2012). Big data analytics in biomedical research. *Biomedical Computation Review*, 2, 14–21.
- Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), 51–59.
- Wactlar, H., Pavel, M., & Barkis, W. (2011). Can computer science save healthcare? *IEEE Intelligent Systems*, 26(5), 79–83.

Data Scientist

Derek Doran

Department of Computer Science and
Engineering, Wright State University, Dayton,
OH, USA

Synonyms

Data analyst; Data analytics; Data hacker;
Statistician

Definition/Introduction

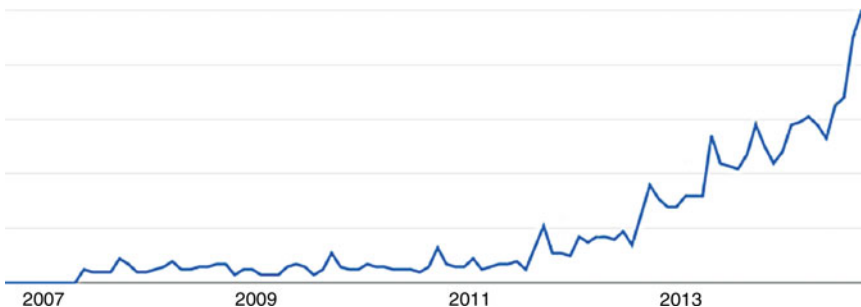
A “Data Scientist” is broadly defined as a professional that systematically performs operations over data to acquire knowledge or discover non-obvious trends and insights. They are employed by organizations to acquire such knowledge and trends from data using sophisticated computational systems, algorithms, and statistical techniques. Given the ambiguity of how a data scientist extracts knowledge and how the data she operates on may be defined, the term does not have a narrow but universally applicable definition. Data scientists use computational tools and computer programming skills, their intellectual foundation in mathematics and statistics, and at-hand domain knowledge to collect, deconstruct, and fuse data from (possibly) many sources, compare models, visualize, and report in non-technical terms new insights that routine

data analysis methods cannot reveal. The demands for data scientists are high, and the field is projected to see continued growth over time. Leading universities now offer undergraduate degrees, graduate degrees, and graduate certificates in data science.

Defining a “Data Scientist”

If a data scientist is one who transforms and applies operations over data to acquire knowledge, nearly any individual processing, analyzing, and interpreting data may be considered to be one. For example, a zoologist that records eating habits of animals, a doctor who reviews a patient’s history of blood pressure, a business analyst that summarizes data in an Excel spreadsheet, and a school teacher who computes final grades for a class are data scientists in the broadest sense. It is for this reason that the precise definition of what a data scientist is, and the skills necessary to fulfill the position, may be a controversial topic. Spirited public debate about what constitutes one to have the job title “data scientists” can be seen on professional discussion boards and social networks across the Web and in professional societies. The meteoric rise in popularity of this term, as identified by Google Trends in Fig. 1, leads some to suggest that the popularity of the title is but a trend powered by excitement surrounding the term “Big Data.”

Although the specific definition of the title data scientist varies among organizations, there is agreement about the skills required to fulfill



Data Scientist, Fig. 1 Interest in the term “data scientist” as reported by Google Trends, 2007–2013

the role. Drew Conway’s “data science Venn diagram”, published in 2010, identifies these agreed upon characteristics of a data scientist. It includes: (i) hacking skills, i.e., the ability to use general purpose computational tools, programming languages, and system administration commands to collect, organize, divide, process, and run data analysis across modern computing platforms; (ii) mathematics & statistics knowledge, which encompasses the theoretical knowledge to understand, choose, and even devise mathematical and statistical models that can extract complex information from data; and (iii) substantive expertise about the domain that a data set has come from and/or about the *type* of the data being analyzed (e.g., network, natural language, streaming, social, etc.) so that the information extracted can be meaningfully interpreted.

Data Hacking and Processing

Data scientists are often charged with the process of collecting, cleaning, dividing, and combining raw data from a number of sources prior to analysis. The raw data may come in a highly structured form such as the result of relational database queries, comma or tab-delimited files, and files formatted by a data interchange language including xml or json. The raw data may also carry a semistructured format through a document markup language such as html, where markup tags and their attributes suggest a document structure but the content of each tag is unstructured. Datasets may even be fully unstructured, formatted in ways such as audio recordings, chat transcripts, product reviews written in natural language, books, medical records, or analog data. **Data wrangling** is thus an important data hacking process where datasets of any form are collected and then transformed or mapped into a common, structured format for analysis. Leading open source data wrangling tools include **OpenRefine** and the **Pandas** package for the **Python programming language**. Data scientists will also turn to Linux shell commands and scripts to maximize their ability to collect (e.g., sed, jq, scrape) and transform raw data into alternative formats (cut, awk, grep, cat, join). In a second

process called **data fusion**, data from different, possibly heterogeneous, sources are melded into a common format and then joined together to form a single set of data for analysis.

Data scientists perform their analysis using advanced computational frameworks built on high performance, distributed computing clusters that may be offered by cloud services. They thus have working knowledge about popular frameworks deployed in industry, such as **Hadoop** or **Spark** for large-scale batch processing of data and **Storm** for real-time data analysis. They also may know how to build, store, and query data in **SQL** relational database systems like MySQL, MSSQL, and Oracle, as well as less traditional **noSQL** database management systems, including HBase, MongoDB, Oracle NoSQL, CouchDB, and Neo4j, which emphasize speed and flexibility of data representation over data consistency and transaction management. In both “small” and “big” data settings, data scientists often utilize statistical programs and packages to build and run A/B testing, machine learning algorithms, deep learning systems, genetic algorithms, natural language processing, signal processing, image processing, manifold learning, data visualization, time series analysis, and simulations. Towards this end, they often have working knowledge of a statistical computing software environment and programming language. **R** or **Python** are often selected because of their support for a number of freely available, powerful packages for data analytics.

Mathematical and Statistical Background

Data scientists may be asked to analyze high-dimensional data or data representing processes that change over time. They may also be charged with making predictions about the future by fitting complex models to data and execute data transformations. Techniques achieving these tasks are rooted in the mathematics of calculus, linear algebra, and probability theory, as well as statistical methods. Calculus is used by data scientists in a variety of contexts but most often to solve model optimization problems. For example, data scientists often devise models that relate data attributes to a desired outcome and include parameters

whose values cannot be estimated from data. In these scenarios, analytical or computational methods for identifying the value of a model parameter “best explaining” observed data take derivatives find the direction and magnitude of parameter updates that reduce a “loss” or “cost” function. Advanced machine learning models involving **neural networks** or **deep learning systems** require a background in calculus to evaluate the backpropagation learning algorithm.

Data scientists imagine data that carry n attributes as an n -dimensional vector oriented in an n -dimensional vector space. Linear algebraic methods are thus used to project, simplify, combine, and analyze data through geometric transformations and manipulations of such vectors. For example, data scientists use linear algebraic methods to simplify or eliminate irrelevant data attributes by projecting vectors representing data into a lower dimensional space. Many statistical techniques and machine learning algorithms also rely on the spectrum, or the collection of eigenvalues, of matrices whose rows are data vectors. Finally, data scientists exploit linear algebraic representations of data in order to build computationally efficient algorithms that operate over data.

Data scientists often use simulation in order to explore the effect of different parameter values to a desired outcome and to test whether or not a complex effect observed in data may have arisen simply by chance. Building simulations that accurately reflect the nature of a system from which data has come from require a data scientist to codify its qualities or characteristics probabilistically. They will thus fit single or multivariate discrete and continuous probability distributions to the data and may build Bayesian models that reflect the conditionality’s latent within the system being simulated. Data scientists thus have a sound understanding of probability theory and probabilistic methods to create accurate models and simulations they draw important conclusions from.

Statistical analyses that summarize data, identify its important factors, and yield predictive analytics are routinely used by data scientists. Data summarizations are performed using not only summary statistics (e.g. mean, median, and

mode of data attributes) but also by studying characteristics about the distribution of the data (variance, skew, and heavy-tailed properties). Depending on the nature of the data being studied, relevant factors may be identified through regression, mixed effect, or unsupervised machine learning methods. Predictive analytics are also powered by machine learning algorithms, which are chosen or may even be developed by a data scientist based on the statistical qualities of the data.

Domain Expertise

Data scientists are also equipped with domain- and organization-specific knowledge in order to translate their analysis results into actionable insights. For example, data scientists evaluating biomedical or biological data have some training in the biological sciences, and a data scientist that studies interactions among individuals has training in social network analysis and sociological theory. Once employed by an organization, data scientists immediately begin to accrue organization-specific knowledge about the company, the important questions they need their analysis to answer, and the best method for presenting their results in a nontechnical fashion to the organization. Data scientists are unable to create insights from an analysis without sufficient domain-specific expertise and cannot generate value or communicate their insights to a company without organization-specific knowledge.

The Demand for Data Scientists

For all but the largest and influential organizations, identifying and recruiting an individual with strong theoretical foundations, familiarity with state-of-the-art data processing systems, ability to hack at unstructured data files, an intrinsic sense of curiosity, keen investigative skills, and an ability to quickly acquire domain knowledge is a tremendous challenge. Well-known consulting and market research group McKinsey Global Institute project deep supply and demand gaps of analytic talent across the world, as defined by talent having knowledge of probability, statistics, and machine learning. For example, McKinsey projects that by the end of 2017, the United States

will face a country-wide shortage of over 190,000 data scientists, and of over 1.5 million managers and analysts able to lead an analytics team and are able to interpret and act on the insights data scientists discover.

Data Science Training

To address market demand for data scientists, universities and institutes around the world now offer undergraduate and graduate degrees as well as professional certifications in data science. These degrees are available from leading institutions including Harvard, University of Washington, University of California Irvine, Stanford, Columbia, Indiana University, Northwestern, and Northeastern University. Courses within these certificate and Master's programs are often available online and in innovative formats, including through massive open online courses.

Conclusion

Despite the challenge of finding individuals with deep mathematical, computational, and domain specific backgrounds, the importance for organizations to identify and hire well-trained data scientists has never been so high. Data scientists will only continue to rise in value and in demand as our global society marches forward towards an ever more data-driven world.

Cross-References

- [Big Variety Data](#)
- [Computer Science](#)
- [Computational Social Sciences](#)
- [Digital Storytelling, Big Data Storytelling](#)
- [Mathematics](#)
- [R-Programming](#)
- [Statistics](#)

Further Reading

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review*, 90, 70–76.

Granville, V. (2013). Data science programs and training currently available. *Data Science Central*. Data Science Central. Web. Accessed 04 Dec 2014.

Kandel, S., et al. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4), 271–288.

Lund, S. (2013). *Game changers: Five opportunities for US growth and renewal*. McKinsey Global Institute.

Walker, D., & Fung, K. (2013). Big data and big business: Should statisticians join in? *Significance*, 10(4), 20–25.

Data Security

- [Data Provenance](#)

Data Security Management

- [Data Provenance](#)

Data Service

- [Data Repository](#)

Data Sharing

Tao Wen

Earth and Environmental Systems Institute,
Pennsylvania State University, University Park,
PA, USA

Definition

In general, data sharing refers to the process of making data accessible to data users. It often happens through community-specific or general data repositories, personal and institutional websites, and/or data publications. A data repository is a place storing data and providing access to users. Data sharing is particularly encouraged in research communities although the extent to which data are

being shared varies across scientific disciplines. Data sharing links data providers and users, and it benefits both parties through improving the reproducibility and visibility of research as well as promoting collaboration and fostering new science ideas. In particular, in the big data era, data sharing is particularly important as it makes big data research feasible by providing the essential constituent – data. To ensure effective data sharing, data providers should follow findability, accessibility, interoperability, and reusability (FAIR) principles (Wilkinson et al. 2016) throughout all stages of data management, a broader topic underpinned by data sharing.

FAIR Principles

Wilkinson et al. (2016) provide guidelines to help the research community to improve the findability, accessibility, interoperability, and reusability of scientific data. Based on FAIR principles, scientific data should be transformed into a machine-readable format, which becomes particularly important given that an enormous volume of data is being produced at an extremely high velocity. Among those four characteristics of FAIR data, reusability is the ultimate goal and the most rewarding step.

Findability

Data sharing starts with making the data findable to users. Both data and metadata should be made available. Metadata are used to provide information about one or more aspects of the data, e.g., who collect the data, the date/time of data collection, and topics of collected data. Each dataset should be registered and assigned a unique identifier such as a digital object identifier (DOI). Each DOI is a link redirecting data users to a webpage including the description and access of the associated dataset. Both data and metadata should be formatted following formal, open access, and widely endorsed data reporting standard (e.g., [schema.org: https://schema.org/Dataset](https://schema.org/Dataset)). Those datasets fulfilling these standards can be cataloged by emerging tools for searching datasets (e.g.,

Google Dataset Search: <https://toolbox.google.com/datasetsearch>). Currently, it is more common that data users will search for desired datasets through discipline-specific data repositories (e.g., EarthChem: <https://www.earthchem.org/> in earth sciences).

Accessibility

Both data and metadata should be provided and can be transferred to data users through data repository. Broadly speaking, data repository can be personal- or institutional-level websites (e.g., Data Commons at Pennsylvania State University: <http://www.datacommons.psu.edu>) and discipline-specific or general databases (e.g., EarthChem). Data users should be able to use the unique identifier (e.g., DOI) to locate and access a dataset.

Interoperability

As more interdisciplinary projects are proposed and funded, shared data from two or more disciplines often need to be integrated for data visualization and analysis. To achieve interoperability, data and metadata should not only follow broadly adopted reporting standards but also use vocabularies to further formalize reported data. These vocabularies should also follow FAIR principles. The other way to improve interoperability is that data repositories should be designed to provide shared data in multiple formats, e.g., CSV and JavaScript Object Notation (JSON).

Reusability

Enabling data users to reuse shared data is the ultimate goal. Reusability is the natural outcome if data (and metadata) to be shared meet the rules mentioned above. Shared data can be reused for testing new science ideas or for reproducing published results along with the shared data.

The Rise of Data Sharing

Before the computer age, it was not uncommon that research data were published and deposited as paper copies. Transferring data to users often

required individual request sent to the data provider. The development of the Internet connects everyone and allows data sharing almost in real time (Popkin 2019).

Nowadays more data are shared through a variety of data repositories providing access to data users. The scientific community including funders, publishers, and research institutions has started to promote the culture of data sharing and making data open access. For example, the National Science Foundation requires data management plans in which awardees need to describe how research data will be stored, published, and disseminated. Many publishers, like *Springer Nature*, also require authors to deposit their data in general or discipline-specific data repository. In addition to sharing data in larger data repositories funded by national or international agencies, many research institutions start to format and share their data in university-sponsored data repositories for the purpose of long-term data access.

In some disciplines, for example, Astronomy and Meteorology, where data collection often relies on large and expensive facilities (e.g., satellite, telescope, a network of monitoring stations) and the size of dataset is often larger than what one research group can analyze, data sharing is a common practice (Popkin 2019). In some other disciplines, some researchers might be reluctant to share data for varying reasons. These reasons can be in the processes of data publication and data citation. Some of these reasons include:

- (1) Researchers are concerned that they might get scooped if they share data too early.
- (2) Researchers might lack of essential expertise to format their data to certain standard.
- (3) Funding that supports data sharing might not be available to these researchers to pay for their time to make data FAIR.
- (4) The support for building data repositories is insufficient in some disciplines.
- (5) The research community fails to treat data sharing as important as publishing journal article.

- (6) Insufficient credit has been given to data providers as data citation might not be done appropriately by data users.

To address some of these problems, all stakeholders of data sharing are working collaboratively. For example, European Union projects FOSTER Plus and OpenAIRE provide training opportunities to researchers on open data and data sharing. The emerging data journals, e.g., *Nature Scientific Data*, provide a platform for researchers to publish and share their data along with descriptions. Many funders, including the National Science Foundation, have allowed repository fees on grants (Popkin 2019).

Best Practices

The United States National Academies of Sciences, Engineering, and Medicine published a report in 2018 (United States National Academies of Sciences, Engineering, and Medicine 2018) to introduce the concept of *Open Science by Design*, in which a series of improvements were recommended to be implemented throughout the entire research life cycle to ensure open science and open data. To facilitate data sharing and to promote open science, some initiatives listed below were recommended:

Data Generation

During data generation, researchers should consider collecting data in a digital form other than noting down data on a paper copy, e.g., a laboratory notebook. Many researchers are now collecting data in electronic forms (e.g., comma-separated values or CSV files). In addition, researchers should use tools compatible with open data, and also adopt automated workflows to format and curate generated data. These actions taken at the early stage of research life cycle can help avoid many problems in data sharing later on.

Data Sharing

After finishing preparing data, researchers should pick one or more data repositories to share their data. Data to be shared include not only data but also metadata and more. For example, World Data System (2015) recommended that data, metadata, products, and information produced from research all should be shared although national or international jurisdictional laws and policies might apply. Researchers should consult with funders or publishers about recommended data repositories into which they can deposit data. One example list of widely used data repositories (both general and discipline-specific) can be found here: <https://www.nature.com/sdata/policies/repositories>.

Conclusion

Data sharing act as a bridge linking both data providers and users, and it is particularly encouraged in the research community. Data sharing can benefit the research community in many ways including (1) improving the reproducibility and visibility of research, (2) promoting collaboration, and inspiring new science ideas, and (3) shared data can be used as a vehicle to foster the communication between academia, industry, and the general public (e.g., Brantley et al. 2018). To facilitate effective data sharing, researchers should follow FAIR principles (findability, accessibility, interoperability, and reusability) when they generate, format, curate, and share data.

Cross-References

► [Data Repository](#)

Further Reading

- Brantley, S. L., Vidic, R. D., Brasier, K., Yoxtheimer, D., Pollak, J., Wilderman, C., & Wen, T. (2018). Engaging over data on fracking and water quality. *Science*, 359 (6374), 395–397.
- Popkin, G. (2019). Data sharing and how it can benefit your scientific career. *Nature*, 569(7756), 445.

United States National Academies of Sciences, Engineering, and Medicine. (2018). Open science by design: Realizing a vision for twenty-first century research. National Academies Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

World Data System. (2015). World Data System (WDS) Data Sharing Principles. Retrieved 22 Aug 2019, from <https://www.icsu-wds.org/services/data-sharing-principles>.

Data Storage

Omar Alghushairy and Xiaogang Ma
Department of Computer Science, University of Idaho, Moscow, ID, USA

Synonyms

[Data science](#); [FAIR data](#); [Storage media](#); [Storage system](#)

Introduction

Data storage means storing and archiving data in electronic storage devices that are dedicated for preservation, where data can be accessed and used at any time. Storage devices are hardware that are used for reading and writing data through a storage medium. Storage media are physical materials for storing and retrieving data. The popular data storage devices are hard drives, flash drives, and cloud storage. The term *big data* reflects not only the massive volume of data but also the increased velocity and variety in data generation and collection, for example, the massive amounts of digital photos shared on the Web, social media networks, and even the Web search records. Many conventional documents such as books, newspapers, and blogs can also be data sources in the digital world. Storing big data in appropriate ways will greatly support data discovery, access, and analytics. Accordingly, various data storage devices and technologies have been developed to increase the efficiency

in data management and enable information extraction and knowledge discovery from data. In domain-specific fields, a lot of scientific researches has been conducted to tackle the specific requirements on data collection, data formatting, and data storage, which have also generated beneficial feedback to computer science.

Data storage is a key step in the data science life cycle. At the early stage of the cycle, well-organized data will provide strong support to the research program where the data is collected. At the late stage, the data can be shared and made persistently reusable to other users. The FAIR data principles (findable, accessible, interoperable, and reusable) provide a set of guidelines for data storage.

Data Storage Devices

There are many different types of devices that store data in digital forms, which have the fundamental capacity measurement unit called bit, and every eight bits are equal to one byte. Often, the data storage device is measured in megabytes (MB), gigabytes (GB), terabytes (TB), and other bigger units. The data storage devices are categorized into two types based on their characteristics: the primary storage and the secondary storage.

Primary storage devices such as cache memory, random-access memory (RAM), and read-only memory (ROM) are connected to a central processing unit (CPU) that reads and executes instructions and data stored on them. Cache memory is very fast memory, which is used as the buffer between the CPU and RAM. RAM is temporary memory, which means the content of stored data is lost once the power is turned off. ROM is nonvolatile memory, so the data stored on it cannot be changed because it has become permanent data (Savage and Vogel 2014). In general, these memories have limited capacity, where it is difficult to handle big data streaming.

Secondary storage such as hard disk drive (HDD), solid-state drive (SSD), server, CD, and DVD are external data storage that are not

connected to the central processing unit (Savage and Vogel 2014). This type of data storage device is usually used to increase computer capacity. Secondary storage is nonvolatile, and the data can be retained. HDD stores the data in the magnetic platter, and it uses the mechanical spindle to read and write data. The operating system identifies the paths and sectors of data stored on the platters. SSD is faster than HDD because it is a flash drive, which stores data in microchips and has no mechanical parts. Also, SSD is smaller in size, is less in weight, and is energy-efficient in comparison with HDD.

Technologies

In recent years, data has grown fast and has become massive. With so many data generating sources, there is an urgent need to provide technologies that can deal with the storage of big data. This section provides an overview of well-known data storage technologies that are able to manipulate large volumes of data, such as relational database, NoSQL database, distributed file systems, and cloud storage.

Relational Database: The relational system that emerged in the 1980s is described as a cluster of relationships, each relationship having a unique single name. These relationships interconnect a number of tables. Each table contains a set of rows (records) and columns (attributes). The set of columns in each table is fixed, and each column has a specific pattern that is allowed to be used. In each row, the record represents a relationship that linked a set of values together. Relational database is functional in data storage, but it also has some limitations that make it less efficient to deal with big data. For example, relational database cannot tackle unstructured data. For datasets with network or graph patterns, it is difficult to use relational database to find the shortest route between two data points.

NoSQL Database: “Not only SQL” (NoSQL) database is considered the most important big data storage technology in database management systems. It is a method that depends on disposal of restrictions. NoSQL databases aim to eliminate

complex relationships and provide many ways to preserve and work on data for specific use cases, such as storing full-text documents. In NoSQL database, it is not necessary for data elements to have the same structure, because it is able to deal with structured, unstructured, and semi-structured data (Strohbach et al. 2016).

Distributed File Systems (DFS): DFS manages datasets that are stored in different servers. Moreover, DFS accesses the datasets and processes them as if they are stored in one server or device. The Hadoop Distributed File System (HDFS) is the most popular method in the field. HDFS separates the data into multiple servers. Thus, it supports big data storage and high-efficiency parallel processing (Minelli et al. 2013).

Cloud Storage: Cloud storage can be defined as servers that contain large storage space where users can manage their files. In general, this service is provided by companies known in the field of cloud storage. Cloud storage led to the term cloud computing, which means using applications over a virtual interface by connecting to the Internet. For example, Microsoft installs the Microsoft Office on its cloud servers. If a user has an account in the Microsoft cloud storage service and an available Internet connection through a computer or smart phone, the user will be allowed to use the cloud system by logging into the account from anywhere. Besides cloud computing, cloud storage also has many other features, such as file synchronization, file sharing, and collaborative file editing.

Impacts of Big Data Storage

Based on a McKinsey Global Institute study, the information that has been captured through organizations about their customers, operations, and suppliers by digital systems has been estimated as trillions of bytes. That means data volume grows at a great rate, so it needs advanced tools and technologies for storing and processing. Data storage has played a major role in the big data revolution (Minelli et al. 2013).

Many companies are using emotion and behavior analysis from their data or social media to identify their audiences and costumers to predict the marketing and sales results. Smart decisions reduce costs and improve productivity. Data is the basis for informed big business decision-making. Analyzing the data offers more information options to make the right choice.

There are many techniques for managing the big data, but Hadoop is currently the best technology for this purpose. Hadoop offers the data scientists and data analysts flexibility to deal with data and extract information from it whether the data is structured or unstructured, and it offers many other convenient services. Hadoop is designed to follow up on any system failures. It constantly monitors the stored data on the server. As a result, Hadoop provides reliable, fault-tolerant, and scalable servers to store and analyze data at a low cost.

The development of cloud storage with the widespread use of Internet services, as well as the development of mobile devices such as smart phones and tablets, has enhanced the spread of cloud storage services. Many people carry their laptops when they are not in their offices, and they can easily access their files through their own cloud storage over the Internet. They can use cloud storage services like Google Docs, Dropbox, and many more to access their files wherever they are and whenever they want.

Companies are increasingly using cloud storage for several reasons, most notably because cloud services are becoming cheaper, faster, and easier to maintain and retrieve data. In fact, cloud storage is the better option for a lot of companies to address challenges caused by the lack of office space, the inability to host servers, and the expensive cost of using servers in the company, in terms of maintenance and cost of purchase. By using cloud storage, companies can save the servers' space and cost for other things. Google, Amazon, and Microsoft are the most popular companies in cloud storage services, just to name a few.

Structured, Unstructured, and Semi-structured Data

There are various forms of data that are stored, such as texts, numbers, videos, etc. These data can be divided into the following three categories: structured, unstructured, and semi-structured data. Structured data is considered high-level data that is in an organized form, such as data in an Excel sheet. For example, a university database has around half a million pieces of information for about 20 thousand students, which contain names, phone numbers, addresses, majors, and other data. Unstructured data is random and disorganized data, for example, data that is presented on a social network, such as text and multimedia data. Various unstructured data are posted to social media platforms like Twitter and YouTube every day. Semi-structured data is provided by several types of data combined to represent the data in a specific pattern or structure. For example, information about a user's call contains an entity of information based on the logs of the call center. However, not all the data is structured, such as a complaint recorded in audio format, which is unstructured, so it is hard to be synthesized in data storage (Minelli et al. 2013).

FAIR Data

FAIR data is a new point of view on data management, which follows the guidelines of findability, accessibility, interoperability, and reusability (Wilkinson et al. 2016). FAIR data focuses on two principles which enhance machine's ability for finding and using data automatically and supporting the data reuse via humans.

Findability is based on placing the data with its metadata in searchable and global identifiers; then looking for data through several links on the World Wide Web should be possible. Accessibility is based on ensuring easy access to data and its metadata (Ma 2019) through the Internet by an authorized person or a machine.

Metadata should be made accessible even if the data is not accessible. Interoperability is based on containing qualified references for both data and metadata and by representing the records in formal, shareable, and machine-readable language. Reusability is based on detailed information of metadata with accessible license for suitable citation to the data. In addition, software tools and other related provenance information should also be accessible to support data reuse.

Cross-References

► [Data Center](#)

Further Reading

- Ma, X. (2019). Metadata. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data*. Cham: Springer. https://doi.org/10.1007/978-3-319-32001-4_135-1.
- Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big data, big analytics: Emerging business intelligence and analytic trends for today's businesses*. Hoboken: Wiley.
- Savage, T. M., & Vogel, K. E. (2014). An introduction to digital multimedia (2nd ed.). Jones & Bartlett Learning Publication, Burlington, MA, USA.
- Strohbach, M., Daubert, J., Ravkin, H., & Lischka, M. (2016). Big data storage. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy*. Cham: Springer.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

Data Store

► [Data Repository](#)

Data Stream

► [Data Streaming](#)

Data Streaming

Raed Alsini and Xiaogang Ma
Department of Computer Science,
University of Idaho, Moscow, ID, USA

Synonyms

Data science; Data stream; Internet of Things (IoT); Stream reasoning

Introduction

Data has become an essential component of not only research but also of our daily life. In the digital world, people are able to use various types of technology to collect and transmit big data, which has the features of overwhelming volume, velocity, variety, value, and veracity. More importantly, big data represents a vast amount of information and knowledge to be discovered. The Internet of Things (IoT) is interconnected with big data. IoT applications use data stream as a primary way for data transmission and make data stream a unique type of big data. A data stream is a sequence of data blocks being transmitted. The real-time feature of the data stream requires corresponding technologies for efficient data processing. Streaming the data is built upon resources that are commonly used for communication, web activity, E-commerce, and social media. How the data is processed determines how information can be extracted from the data stream. Analyzing the data stream through queries ensures and improves the efficiency on data under the aspect of data science. Many techniques can be used in data stream processing, among which data mining is the most common approach used for detecting data latency, frequent pattern, and anomaly values, as well as for classification and clustering. The computer science community has created many open-source libraries for the data stream and has built various best practices to facilitate the applications of data stream in different disciplines.

The Internet of Things (IoT)

IoT is a crucial source of big data. It consists of connected devices in a network system. Within IoT, as a part of the data stream application, many tools are already widely used, such as radio-frequency identification (RFID). For example, a production company can use RFID on their system to track products such as automobiles, toys, and clothes. By doing so, the workers in that company could monitor and resolve issues during the production process. Other examples can be seen in the data resulted from chip devices. A smartphone, with many sensor chips embedded, is able to measure and record its user's activities, such as location, footsteps, heart rate, and calorie, just to name a few.

Sensors are essential components of IoT. They generate the data stream and transmit them into many applications, such as those estimating weather conditions. For instance, historical and live weather records of rain, snow, wind, and others are input into a weather analysis system to generate hourly, daily, and weekly predictions. Automotive industries also equip many sensors in vehicles to help reduce potential traffic accidents. For example, the distance detection radar is a common component of many automobiles nowadays. It can detect the distance and space between the automobile and a pedestrian or a building to prevent any injury when the distance approaches a certain minimum value.

Data Science Aspect in Data Stream

In recent years, data is the “crude oil” to drive technological and economic development. There is an extremely high demand, and almost everyone uses it. We need to refine crude oil before using it. It is the same with data. We can benefit from the data only when data processing, mining, analysis and extraction are able to provide useful results.

Using the data stream in data science involves understanding the data life cycle. Usually, it begins with collecting the data from the

sources. Nowadays, data stream collection can be seen on search engines, social media, IoT, and marketing. For instance, the Google Trends generates a massive amount of data by searching certain topics on the Web. After that, it can provide the result based on what a user is looking for with a specific range of time. The benefit of processing the data stream is getting the right information immediately. The processing needs methods and models. Two common standard models are batch processing and stream processing.

Batch processing can handle a large amount of data by first collecting the data over time and then doing the processing. For example, the operating system on a computer can optimize the sequencing of jobs to make efficient usage of the system. Micro-batch is a modified model of batch processing. It groups data and tasks into small batches. Completing the processing of a batch on this model is based on how the next batch is received.

Stream processing is a model used for processing the data without waiting for the next data to arrive. The benefit of stream processing is that the system can receive the data quickly. For example, an online banking application runs stream processing when a customer buys a product. Then the bank transaction is verified and executed without fail. Stream processing can handle a huge amount of data without suffering any issues related to data latency. A sensor network that generates massive data can be organized easily under this method.

Many technologies can be used to store the data stream in a data life cycle. Amazon Web Services (AWS) provide several types of tools to support the various needs in storing and analyzing data. Apache Spark is an open-source use for cluster computing framework. Spark Streaming uses the fast scheduling capability of Apache Spark to implement streaming analytics. It groups data in micro-batches and makes transformations on them. Hadoop tool, which is another platform under Apache, uses the batch processing to store massive amounts of data. It sets up a framework for distributed data processing by using the MapReduce model. Both Spark and Hadoop

have several libraries that can deal with many types of data stream.

Data mining as a part of data science is used to discover knowledge in data. For data stream mining, it usually involves methods such as machine learning to extract and predict new information. A few widely used methods are clustering, classification, and stream mining on the sensor network.

Clustering is a process of gathering similar data into a group. Clustering deals with unsupervised learning. It means the system does not need to have a label in order to discover the hidden pattern in data. K-means is the most common method used for clustering. Clustering can be used for fraud detection. For example, it is able to find anomaly records on a credit card and inform the card holder.

Classification is a process of identifying the category of a piece of new data. Based on a set of training data, the system can set up several categories and then determine to which category a piece of new data belongs. Classification is one of the supervised learning methods, in which the system learns and determines how to take the right decision. For example, buying and selling holds on the stock market can be done by using this method to make the right decisions based on the data given.

Data Stream Management System (DSMS)

Regardless of how the data stream is preceded and stored, it requires data management in the data life cycle. Managing the data stream can be done using queries as a primary method, such as the Structured Query Language (SQL). SQL is a common language use for managing the database. The data stream management system (DSMS) uses an extended version of SQL known as the Continuous Query Language (CQL). The reason behind CQL is to ensure any continuous data over time can be used on the system.

The operation of CQL can be categorized into three groups: relation-to-relation, stream-to-relation, and relation-to-stream (Garofalakis et al. 2007). Relation-to-relation is usually done under

the SQL query. For instance, the relation between two queries can be expressed using either equal, above, greater, or less symbols. Stream-to-relation is done using the sliding window method. Sliding Window is based on having a window that has historical points when the data is streamed. Specifically, when there are two window sizes, the second window will not begin until the difference between them is removed. Relation-to-stream usually involves the tree method to deal with the continuous query. Detailed operations include insertion, deletion, and relation.

Stream Reasoning

Stream reasoning is about processing the data stream to get a conclusion or decision on continuous information. Stream reasoning handles the continuous information by defining factors on the velocity, volume, and variety of big data. For example, a production company uses several sensors to estimate and predict the types and amounts of raw materials needed for each day. Another example is the detection of fake news on social media. Each social media platform has various users across the world. Streaming reasoning can be used to analyze the features in the language patterns in message spreading.

The semantic web community has proposed several tools that can be used in stream reasoning. The semantic web community introduced RDF for the modeling and encoding of data schemas and ontologies on the fundamental level. The linked open data is an example of how the database can be linked in the semantic web. SPARQL is a query language developed by W3C. SPARQL queries use the triplet pattern of RDF to represent patterns in the data and graph. Recently, the RDF Stream Processing (RSP) working group proposed an extension of both RDF and the SPARQL query to support stream reasoning. For instance, the Continuous SPARQL (C-SPARQL) is an example of the SPARQL language to expand the use of continuous queries.

Practical Approach

In real-world practice, the application of data stream is tied with big data. The current approach data stream usage can be grouped into these categories: scaling data infrastructure, mining heterogeneous information network, graph mining and discovery, and recommender system (Fan and Bifet 2013).

Scaling data infrastructure is about to analyze the data from social media, such as Twitter, which carry various types of data such as video, image, text, or even a hashtag trend. The data generated is based on how the users communicate on a certain topic. It leads to various analytics to understand human behavior and emotion on the communication between users. Snapchat is now another popular social media application that generates and analyzes live data stream based on the location and the event occurred.

Mining heterogeneous information network is about to discover the connections between multiple components such as people, organizations, activities, communication, and system infrastructure. The information network here also includes the relations that can be seen on social networks, sensor networks, graphs, and the Web.

Graphs are being used to represent nodes and their relations, and graph mining is an efficient method to discover knowledge in the big data. For example, Twitter can represent the graph information by visualizing each data type and their relations. Many other graph information can be obtained from the Web. For example, Google has constructed knowledge graphs for various objects and relations.

Recommender system is another approach for analyzing data stream in the big data. Through the collaborative filtering (CF), the queries in a DSMS system can be improved by adding a new CF statement, such as rating. It can extend the functionality of DSMS on finding optimization, query sharing, fragmentation, and distribution. Another strategy is using the content-based model. Several platforms, like Amazon, eBay, YouTube, and Netflix, have already used this in their systems.

Further Reading

- Aggarwal, C. C. (2007). An introduction to data streams. In C. C. Aggarwal (Ed.), *Data streams. Advances in database systems* (Vol. 31). Boston: Springer.
- Dell'Aglío, D., Valle, E. D., Harmelen, F. V., & Bernstein, A. (2017). Stream reasoning: A survey and outlook. *Data Science*, 1–25. <https://doi.org/10.3233/ds-170006>.
- Fan, W., & Bifet, A. (2013). Mining big data. *ACM SIGKDD Explorations Newsletter*, 14(2), 1. <https://doi.org/10.1145/2481244.2481246>.
- Garofalakis, M., Gehrke, J., & Rastogi, R. (2007). *Data stream management: Processing high-speed data streams (Data-centric systems and applications)*. Berlin/Heidelberg: Springer.
- Ma, X. (2017). Visualization. In L. Schintler & C. McNeely (Eds.), *Encyclopedia of Big Data*. Cham: Springer. https://doi.org/10.1007/978-3-319-32001-4_202-1.

Data Synthesis

Ting Zhang

Department of Accounting, Finance and Economics, Merrick School of Business, University of Baltimore, Baltimore, MD, USA

Definition/Introduction

While traditionally data synthesis often refers to descriptive or interpretative narrative and tabulation in studies like meta-analyses, in the big data context, data synthesis refer to the process of creating synthetic data. In the big data context, the digital technology provides unprecedented tremendous data information. The rich data across various fields can jointly offer extensive information about individual persons or organization for finance, economics, health, other research, evaluation, policy making, etc. However, fortunately our laws necessarily protect our privacy and data confidentiality; this necessary data protection becomes increasing important in our big data world where thefts and various levels of data breach could become much easier. The synthetic data has the same or highly similar attributes of the real data for many analytic purposes but masks

the original data for more privacy and confidentiality. Synthetic data was first proposed by Rubin (1993). Data synthesis therefore is a process of replacing identifying, sensitive, or missing values according to multiple imputation techniques based on regression models; the created synthetic data has many of the same statistical properties as the original data (Abowd and Woodcock 2004). Data synthesis includes a *full synthesis* for all variables and all records or a *partial synthesis* for a subset of variables and records.

The Emergence of Data Synthesis

While many statistical agencies disseminate samples of census microdata, the masked public use data sample can be difficult for analysis, either due to limited or even distorted information after masking or due to the limited sample size when multiple data sources merge together. To disguise identifying or sensitive values, agencies sometimes add random noise or use swapping for easy-to-identify at-risk records (Dalenius and Reiss 1982). This introduces measurement error (Yancey et al. 2002). Winkler (2007) showed that synthetic data has the potential to avoid the problems of standard statistical disclosure control methods and has better data utility and lower disclosure risk. Drechsler and Reiter (2010) demonstrate that sampling with synthesis can improve the quality of public use data relative to sampling followed by standard statistical disclosure limitation.

How Data Synthesis Is Conducted?

For synthetic data, sequential regression imputation is used, with often a regression model for a given variable to impute and replace the values. The process is then repeated for other variables. Specifically, according to Drechsler and Reiter (2010), the data agency typically follows the following four steps:

- (i) Selects the set of values to replace with imputations

- (ii) Determines the synthesis models for the entire dataset to take use of all available information
- (iii) Repeatedly simulates replacement values for the selected data to create multiple, disclosure-protected populations
- (iv) Releases samples from the populations

The newly created samples have a mixture of genuine and simulated data.

Main Challenges

For synthetic data, the challenges include using appropriate inferential methods for different process of data generation. The algorithm could result in different synthetic datasets for different orderings of the variables or possibly different orderings of the records. However, it is not anticipated to affect data utility (Drechsler and Reiter 2010). Essentially, synthetic data only replicates certain specific attributes or simulate general trends of the data; they are not exactly the same as the original dataset for all purposes.

Application of Synthetic Data

Synthetic data is typically used among national statistics agencies because of the potential advantages of synthetic data over data with standard disclosure limitation, such as the American Community Survey, the Survey of Income and Program Participation, the Survey of Consumer Finances, and the Longitudinal Employer Household Dynamics Program. Synthetic data from other national agency experience outside the United States include German Institute for Employment Research and Statistics New Zealand.

Abowd and Lane (2004) describe an ongoing effort, called “Virtual Research Data Centers” or “Virtual RDC”, to benefit both the research community and statistical agencies. In “Virtual RDC,” multiple public use synthetic data sets can be created from a single underlying confidential file and customized for different uses. It is called “Virtual RDC” because the synthetic data is maintained at a

remote site outside with the same computer environment as at the agency’s restricted access Research Data Centers. Researchers can access the synthetic files at the Virtual RDC. The virtual RDC is now operational at the Cornell University.

In addition to government agencies, synthetic data can also be used to create research samples of confidential administrative data or, as a technique, to impute survey data. For example, the national longitudinal Health and Retirement Study that collects survey data on American older adults biannually. Also, the Ewing Marion Foundation has been collecting annual national Kauffman Survey Data longitudinally for years. Both survey data sets have certain components of data synthesis used to impute the data.

Conclusion

In the context of big data, data synthesis refers to the process of creating synthetic data. The limitation of masked public use data makes data synthesis particularly valuable. Data synthesis offers the necessary rich data information needed for numerous research and analysis without sacrificing data privacy and confidentiality in the big data world. Synthetic data adopts regression-based multiple imputations to replace identifying, sensitive, or missing values, which helps to avoid standard statistical disclosure control methods in public use government data samples. With some challenges, synthetic data now is widely used across government or other data agencies and can be used for other data purposes, including imputing survey data, as well.

References

- Abowd, J.M., & Lane, J.I. (2004). New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. In Domingo-Ferrer, J. & Torra, V. (Eds.), *Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004, Barcelona, Spain, June 9-11, 2004, Proceedings*. (pp. 282-289). Berlin: Springer.
- Abowd, J. M., & Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In Domingo-Ferrer, J. & Torra, V.

- (Eds.), *Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004, Barcelona, Spain, June 9-11, 2004, Proceedings*. (pp. 290–297). Berlin: Springer.
- Dalenius, T., & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73–85.
- Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492), 1347–1357.
- Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462–468.
- Winkler, W. E. (2007). *Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified*. Tech. Rep., U.S. Census Bureau Research Report Series, No. 2007–21.
- Yancey, W. E., Winkler, W. E., & Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer (Ed.), *Inference control in statistical databases* (pp. 135–152). Berlin: Springer.

Data Tidying

► Data Cleansing

Data Virtualization

Gagan Agrawal
School of Computer and Cyber Sciences, Augusta
University, Augusta, GA, USA

Data Virtualization is the ability to support a virtual (more abstract) view of a complex dataset. Involving several systems built to support relational views on complex array datasets, automatic data virtualization was introduced in 2004 (Weng et al. 2004). The motivation was that scientific datasets are typically stored as binary or character flat-files. Such low-level layouts enable compact storage and efficient processing, but they make the specification of processing much harder.

In view of this, there recently has been increasing interest in data virtualization, and data services to support such virtualization. Based on the virtualization, low-level, compact, and/or

specialized data formats can be hidden from the applications analyzing datasets. However, supporting it can require significant effort. For each dataset layout and abstract view that is desired, a set of data services need to be implemented. An additional difficulty arises from the fact that the design and implementation of efficient data virtualization and data services oftentimes require interaction of two complementary players. The first player is the scientist who possesses a good understanding of the application, datasets, and their format, but is less knowledgeable about database and data services implementation. The second player is the database developer who is proficient in the tools and techniques for efficient database and data services implementation, but has little knowledge of the specific application.

The two key aspects in the automatic data virtualization approach are as follows.

Designing a Meta-Data Description Language: This description language is expressive enough to present a relational table view for complex multidimensional scientific datasets and describe the low-level data layout. It is very expressive and, particularly, can allow the description as follows:

- Dataset physical layout within the file system of a node.
- Dataset distribution on nodes of one or more clusters.
- The relationship of the dataset to the logical or virtual schema that is desired.
- The index that can be used to make subsetting more efficient.

By using it, the scientist and database developer together can describe the format of the datasets generated and used by the application.

Generating Efficient Data Subsetting and Access Services Automatically: Using a compiler that can parse the meta-data description and generate function code to navigate the datasets, the database developer (or the scientist) can conveniently generate data services that will navigate the datasets. These functions take the user query as input and help create relational tables.

Since this initial work in 2004 (Weng et al. 2004), several other implementations of this approach have been created. The most important ones have involved abstractions on top of NetCDF and HDF5 datasets (Wang et al. 2013; Su and Agrawal 2012).

These implementations addressed some of key challenges in dealing with scientific data. On one hand, this approach does not require data to be loaded into a specific system or to be reformatted. At the same time, it allows the use of a high-level language for specification of processing, which is also independent of the data format. The tool supported SQL select and aggregation queries specified over the *virtual relational table view* of the data. Besides supporting selection over *dimensions*, which is directly supported by HDF5 API also, it also supports queries involving *dimension scales* and those involving *data values*. For this, code for *hyperslab selector* and *content-based filter* was generated in the system. Selection and aggregation queries using novel algorithms also are effectively parallelized.

Implementation has been extensively evaluated with queries of different types, and performance and functionality have been compared against OPeNDAP. Even for subsetting queries that are directly supported in OPeNDAP, the sequential performance of the system is better by at least a factor of 3.9. For other types of queries, where OPeNDAP requires hyperslab selector and/or content-based filter code to be written manually, the performance difference is even larger. In addition, the system is capable of scaling performance by parallelizing the queries, and reducing wide area data transfers through server-side data aggregation. In terms of functionality, the system also supported certain state-of-the-art HDF5 features, including dimension scale and compound datatype. A similar implementation was also carried out in context of another very popular format for scientific data, NetCDF (Su and Agrawal 2012).

Since the initial work in this area and also concurrently with more recent implementations in context of HDF5 and NetCDF, a popular development

has been the NoDB approach (Alagiannis et al. 2012). The idea of NoDB is that datasets continue to be stored in raw-files, but are queries using a high-level language. This work is indeed an example of NoDB approach, but distinct in the focus on multidimensional array data.

Further Reading

- Alagiannis, I., Borovica, R., Branco, M., Idreos, S., & Ailamaki, A. (2012, May). NoDB: Efficient query execution on raw data files. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data* (pp. 241–252).
- Su, Y., & Agrawal, G. (2012, May). Supporting user-defined subsetting and aggregation over parallel netcdf datasets. In *2012 12th IEEE/ACM international symposium on cluster, cloud and grid computing (ccgrid 2012)* (pp. 212–219). IEEE.
- Wang, Y., Su, Y., & Agrawal, G. (2013, May). Supporting a light-weight data management layer over hdf5. In *2013 13th IEEE/ACM international symposium on cluster, cloud, and grid computing* (pp. 335–342). IEEE.
- Weng, L., Agrawal, G., Catalyurek, U., Kur, T., Narayanan, S., & Saltz, J. (2004, June). An approach for automatic data virtualization. In *Proceedings. 13th IEEE international symposium on high performance distributed computing, 2004.* (pp. 24–33). IEEE.

Data Visualisation

► Data Visualization

Data Visualization

Jan Lauren Boyles
Greenlee School of Journalism and
Communication, Iowa State University, Ames,
IA, USA

Synonyms

Data visualisation; Dataviz, Datavis; Information visualization; Information visualisation

Definition/Introduction

Data visualization encompasses the planning, production, and circulation of interactive, “graphical representations” that emerge from big data analyses (Ward et al. 2010, 1). Given the volume and complexity of big data, data visualization is employed as a tool to artfully demonstrate underlying patterns and trends – especially for lay audiences who may lack expertise to directly engage with large-scale datasets. Visually depicting such insights thereby broadens the use of big data beyond computational experts, making big data analyses more approachable for a wider segment of society. More specifically, data visualization helps translate big data into the sphere of decision-making, where leaders can more easily integrate insights gleaned from large-scale datasets to help guide their judgments.

Firstly, it is important to fully distinguish data visualization from the manufacture of information graphics (also known as infographics). The most prominent difference rests in the fact that data visualizations (in the vast majority of cases) are constructed with the assistance of computational structures that manage the statistical complexity of the dataset (Lankow et al. 2012; Mauldin 2015). On the other hand, while computer-assisted products are often used to *construct* infographics, the design is often not directly dependent on the dataset itself (Mauldin 2015). Rather, infographic designers often highlight selected statistics that emerge from the dataset, rather than using the entire dataset to fuel the visualization in the aggregate (Lankow et al. 2012; Mauldin 2015). The corpus of data for infographics, in fact, tends to be smaller in scope than the big data outputs of data visualization projects, which typically encompass millions, if not billions, of data points. Additionally, data visualizations are highly interactive to the user (Yau 2012). Such data visualizations also often tether the data to its spatial qualities, particularly emphasizing the interplay of the geographic landscape. A subfield of data visualization – interactive mapping – capitalizes on this feature of large-

scale datasets that contain geolocation (primarily GPS coordinates). Data visualizations are also more likely than infographics to render the output of real-time data streams, rather than providing a snapshot of a dataset as it appeared at one time in the past. Data visualizations, broadly speaking, tend to evolve as the dataset changes over time. To this end, once the data visualization is created, it can quickly incorporate new data points that may emerge in the dynamic dataset.

Data Visualization as Knowledge Translation

Humans have long been naturally inclined toward processing information visually, which helps anchor individual-level decision-making (Cairo 2012). In this context, visualizing complex phenomena predates our current digital era of data visualization (Friendly 2008). The geographic mapping of celestial bodies or the rudimentary sketches of cave wall paintings, for instance, were among the earliest attempts to marry the complexity of the physical world to the abstraction of visual representation (Friendly 2008; Mauldin 2015).

To today’s users, data visualizations can help unpack knowledge so that nonexperts can better understand the topic at hand. As a result, the information becomes more accessible to a wider audience. In relating this knowledge to the general public, data visualization should, ideally, serve two primary purposes. In its optimal form, data visualizations should: (1) contextualize an existing problem by illustrating possible answers and/or (2) highlight facets of a problem that may not be readily visible to a nonspecialist audience (Telea 2015). In this light, data visualization may be a useful tool for simulation or brainstorming. This class of data visualization, or *scientific visualization*, is generally used to envision phenomena in 3D – such as weather processes or biological informatics (Telea 2015). These data visualization products typically depict the phenomena realistically – where the object or

interaction occurs in space. *Informational visualization*, on the other hand, does not prioritize the relationship between the object and space (Telea 2015). Instead, the data visualizations focus upon how elements operate within the large-scale dataset, regardless of their placement or dimension in space. Network maps, designed to characterize relationships between various actors within social confines, would serve as one example of this type of data visualization.

To disseminate information obtained from the visualization publicly, however, the dataviz product must be carefully constructed. First, the data must be carefully and systematically “cleaned” to eliminate any anomalies or errors in the data corpus. This time consuming process often requires reviewing code, ensuring that that ultimate outputs of the visualization are accurate. Once the cleaning process is complete, the designer/developer of the data visualization must be able to fully understand the given large-scale dataset in its entirety and decide which visualization format would fit best with the information needs of the intended audience (Yau 2012). In making trade-offs in constructing the visualization, the large-scale dataset must also be filtered through the worldview of the dataviz’s creator. The choices of the human designer/developer, however, must actively integrate notions of objectivity into the design process, making sure the dataset accurately reflects the entirety of the data corpus. At the same time, the designer/developer must carefully consider any privacy or security issues associated with constructing a public visualization of the findings (Simon 2014).

In the mid-2010s, data visualization has been used selectively, not systematically, within organizations (Simon 2014). In fact, data visualization is not a flawless tool for presenting big data in all cases. In some cases, data visualization may not be the correct tool altogether for conveying the dataset’s findings, if the work can be presented more clearly in narrative or if little variability exists in the data itself (Gray et al. 2012). Taken together, practitioners using data visualization must carefully contemplate every step of the production and consumption process.

When applying data visualization to a shared problem or issue, the ultimate goal is to fully articulate a given problem within the end user’s mind (Simon 2014). Data visualizations may also illustrate the predictive qualities embedded in large-scale datasets. Ideally, data visualization outputs can highlight cyclical behaviors that emerge from the data corpus, helping to better explicate cause and effect for complex societal issues. The data visualization may also help in identifying outliers contained in the dataset, which may be easier to locate graphically than numerically. Through the data visualization, non-specialists can quickly and efficiently look at the output of a big data project, analyze vast amounts of information, and interpret the results. Broadly speaking, data visualizations can provide general audiences a simpler guide to understanding the world’s complexity (Steele and Iliinsky 2010).

The Rise of Visual Culture

Several developments of the digital age have contributed to the broad use of data visualization as part of everyday practice in analyzing big data across numerous industries. Primarily, the rapid expansion of computing power – particularly the emergence of cloud-based systems – has greatly accelerated the ability to process large-scale datasets in real time (Yau 2012). The massive volume of data generated daily in the social space – particularly on social networking sites – has led both academicians and software developers to create new tools to visualize big data created via social streams (Yau 2012). Some of the largest industry players in the social space, such as Facebook and Twitter, have created Application Programming Interfaces (APIs) that provide public access to large-scale datasets produced from user-generated content (UGC) on social networking sites (Yau 2012). Because APIs have proven so popular for developers working in the big data environment, industry-standard file formats have emerged, which enable data visualizations to be more easily created and shared (Simon 2014).

Technical advances in the required tools to physically create visualizations have also made the production of data visualization more open to the lay user. Data visualization does not rely upon a singular tool or computational method; instead, practitioners typically use a wide spectrum of approaches. At one end of the spectrum rests simple spreadsheet programs that can be used by nonspecialists. When paired with off-the-shelf (and often open source) programs – which incorporate a battery of tools to visualize data in scatterplots, tree diagrams and maps – users can create visualizations in hours without skill sets in coding or design. For those with more proficiency with programming, however, tailored and sophisticated data visualizations can be created using Python, D3, or R. Developers are currently pouring significant energies to making these coding languages and scripts even easier for the general public to manipulate. Such advances will further empower non-specialist audiences in the creation of data visualizations.

In the mid-2010s, the movement toward open data has also led to greater public engagement in the production of data visualization. With a nod toward heightened transparency, organizational leaders (particularly those in government) have unlocked access to big data, providing these large-scale datasets to the public for free. Expanded access to big data produced by Western democratic governments, such as open government initiatives in the United States and the United Kingdom, has precipitated the use of data visualization by civic activists (Yau 2012). The creation of data visualizations following from open government initiatives may, in the long term, foster stronger impact measures of public policy, ultimately bolstering the availability and efficiency of governmental services (Simon 2014).

Beyond use by governments, the application of data visualization in industry can help inform the process of corporate decision-making, which can further spur innovation (Steele and Iliinsky 2010). To the busy executive, data visualizations provide an efficient encapsulation of complex datasets that

are too difficult and laborious to understand on their own. A typical use case in the field of business, for instance, centers upon visually depicting return on investment – translating financial data for decision makers who may not fully grasp the technical facets of big data (Steele and Iliinsky 2010). The visualization can also unite decision makers by creating a shared orientation toward a problem, from which leaders can take direct action (Simon 2014). Within the practice of journalism, data visualization is also increasingly accepted as a viable mechanism for storytelling – sharing insights gleaned from big data with the general public (Cairo 2012).

Conclusion

Future cycles of technological change (particularly the entry of augmented and virtual reality) will likely create a stronger climate for visual data, researchers concur. Within the last 20 years alone, the internet has become a more visual medium. And with the rise of the semantic web and the expansion of data streams available, the interconnectivity of data will require more sophisticated tools – such as data visualization – to convey deeper meaning to users (Simon 2014). The entry of visual analytics, for instance, uses data visualizations to drive real-time decision-making (Myatt and Johnson 2011). Despite the expanded use of data visualizations in a variety of use cases, a current knowledge gap of data literacy exists between those creating visualizations and the end users of the products (Ryan 2016). As this gap begins to close, the continued growth of data visualization as a tool to understand the complexity of large-scale datasets will likely broaden public use of big data in future decades.

Cross-References

- [Open Data](#)
- [Social Network Analysis](#)
- [Visualization](#)

Further Reading

- Cairo, A. (2012). *The functional art: An introduction to information graphics and visualization*. Berkeley: New Riders.
- Friendly, M. (2008). A brief history of data visualization. In C. Chen, W. Hardle, & A. Unwin (Eds.), *Handbook of data visualization* (pp. 15–56). Springer: Berlin.
- Gray, J., Chambers, L., & Bounegru, L. (2012). *The data journalism handbook: How journalists can use data to improve the news*. Beijing: O'Reilly Media.
- Lankow, J., Ritchie, J., & Crooks, R. (2012). *Infographics: The power of visual storytelling*. Hoboken: Wiley.
- Mauldin, S. K. (2015). *Data visualizations and infographics*. Lanham: Rowman & Littlefield.
- Myatt, G. J., & Johnson, W. P. (2011). *Making sense of data iii: A practical guide to designing interactive data visualizations*. Hoboken: Wiley.
- Ryan, L. (2016). *The visual imperative: Creating a visual culture of data discovery*. Cambridge: Morgan Kaufmann.
- Simon, P. (2014). *The visual organization: Data visualization, big data, and the quest for better decisions*. Hoboken: Wiley.
- Steele, J., & Iliinsky, N. (2010). *Beautiful visualization: Looking at data through the eyes of experts*. Sebastopol: O'Reilly Media.
- Telea, A. (2015). *Data visualization: Principles and practice*. Boca Raton: Taylor & Francis.
- Ward, M. O., Grinstein, G., & Keim, D. (2010). *Interactive data visualization: Foundations, techniques, and applications*. Boca Raton: CRC Press.
- Yau, N. (2012). *Visualize this: The flowing data guide to design, visualization and statistics*. Indianapolis: Wiley.

Data Visualizations

► Business Intelligence Analytics

Data Warehouse

► Data Mining

Data Wrangling

► Data Cleansing

Database Management Systems (DBMS)

Sandra Geisler¹ and Christoph Quix^{1,2}

¹Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

²Hochschule Niederrhein University of Applied Sciences, Krefeld, Germany

Overview

DBMS have a long history back to the late 1960s starting with navigational or network systems (CODASYL). These were the first to enable managing a set of related data records. In the 1970s, relational systems have been defined by Codd (1970) which are the most important DBMS until today. Relational database systems had the advantage that they could be accessed in a declarative way while navigational systems used a procedural language. As object-oriented programming became more popular in the 1990s, there was also a demand for object-oriented database systems; especially, to meet the requirement of storing more complex objects for engineering, architectural, or geographic applications. However, the idea of object-oriented DBMS as a back-end for object-oriented applications was not completely realized as relational DBMS dominated the market and provided object-relational extensions in the late 1990s.

With the growing popularity of Internet applications around the year 2000, DBMS had to face again new challenges. The semi-structured data model addressed the problems of data heterogeneity, i.e., datasets with objects that have an irregular, hierarchical structure (Abiteboul et al. 1999). The XML data format was used as a representation of semi-structured data which resulted in a demand for XML DBMS. Again, as in the object-oriented case, relational DBMS were equipped with additional functionalities (e.g., XML data type and XPath queries) rather than that XML DBMS became popular.

On the other hand, data was no longer only viewed as a structured container whose

interpretation was left to a human user. More interoperability was required; systems should be able to exchange data not only a syntactical level (as with XML), but also on a semantical level. This led to the need for attaching context and meaning to the data and to create more intelligent applications, which was addressed by the idea of the Semantic Web (Berners-Lee et al. 2001). Linked data aims at providing more semantics by linking information with semantic ontologies or other data items, leading to complex knowledge graphs that need to be managed by graph databases (Heath and Bizer 2011).

By the event of mobile devices, widely available mobile Internet and high bandwidths, the production of data with higher volume, velocity, and variety, challenging requirements for higher scalability, distributed processing, and storage of data came up. Relational systems were no longer able to fulfill these needs appropriately. Data-intensive web applications, such as search engines, had the problem, that many users accessed their services at the same time, posing the same simple queries to the system. Hence, the notion of NoSQL systems came up, which provided very simple but flexible structures to store and retrieve the data while also relaxing the consistency constraints of relational systems. Also, shorter software development cycles required more flexible data formats and corresponding storage solutions, i.e., without a mandatory schema (re-)design step before data can be managed by a DBMS.

NoSQL in most of the cases means non-relational, but many NoSQL systems also provide a relational view of their data and a query language similar to SQL, as many tools for data analysis, data visualization, or reporting rely on a tabular data representation. In the succeeding notions of Not-only-SQL and NewSQL systems, the concepts of NoSQL and relational systems are approaching each other again to combine the advantages of ACID capabilities (ACID is an acronym for desirable properties in transaction management: atomicity, consistency, isolation, and durability (“ACID Transaction” 2009)) of relational systems with the flexibility, scalability, and distributability of NoSQL systems.

Furthermore, due to the availability of cheap sensors, IoT devices, and other data sources producing data at a very high scale and frequency, the need for processing and analyzing data on-the-fly without the overhead and capability of storing it in its entirety was pressing. Hence, specific systems, such as data stream management systems and complex event processing systems, which are able to cope with these data, evolved.

In principle, DBMS can be categorized along various criteria (Elmasri and Navathe 2017). They are distinguished according to the data model they implement (relational, document-oriented, etc.), the number of parallel users (single or multiple user systems), distributability (one node, multiple equal nodes, multiple autonomous nodes), system architecture (client-server, peer-to-peer, standalone. . .), internal storage structures, purpose (graph-oriented, OLTP, OLAP, data streams), cloud-based or not, and many more. Due to space constraints, we are only able to explain some of these aspects in this article.

Architectures

A DBMS architecture is usually described in several layers to have a modular separation of the key functionalities. There are architectures that focus on the internal physical organization of the DBMS. For example, a five-layer architecture for relational DBMS is presented in Härder (2005). It describes in detail the mapping from a logical layer with tables, tuples, and views, to files and blocks. Transaction management is considered as a function that needs to be managed across several layers.

In contrast, the ANSI/SPARC architecture (or three-schema architecture) is an abstract model focusing on the data models related to a database system (Elmasri and Navathe 2017). It defines an internal layer, a conceptual layer, and an external or view layer. The internal layer defines a physical model describing the physical storage and access structures of the database system. The conceptual layer describes the concepts and their attributes, relationships, and constraints which are stored in the database. The external

layer or view layer defines user or application specific views on the data, thus, presenting only a subset of the database. Each layer hides the details of the layers below, which realizes the important principle of data independence. For example, the conceptual layer is independent of the organization of the data at the physical layer.

There are several architecture models according to which DBMS can be implemented. Usually, a DBMS is developed as a client/server architecture where the client and server software are completely separated and communicate via a network or inter-process communication. Another possibility is a centralized system where the DBMS is embedded into the application program. For processing big data, distributed systems with various nodes managing a shard or replica of the database are used. These can be distinguished according to the grade of autonomy of the nodes, e.g., if a master node organizes the querying and storage of data between the nodes or not (peer-to-peer system). Finally, as systems can have now a main memory with more than one TB, in-memory DBMS become popular. An in-memory system manages the data in main memory only and thereby achieves a better performance than disk-based DBMS that always guarantee persistence of committed data. Persistence of data on disk in in-memory DBMS can be also achieved but requires an explicit operation. Examples are SAP HANA, Redis, Apache Derby, or H2.

Transaction Management and ACID Properties

Transaction management is a fundamental feature of DBMS. In the read-write model, a transaction is represented as a sequence of the following abstract operations: read, write, begin (of transaction), commit, and abort. Transaction management controls the execution of transactions to ensure the consistency of the database (i.e., the data satisfies the constraints) and to enable an efficient execution. If the DBMS allows multiple users to access the same data at the same time, several consistency problems may arise (e.g., Dirty-Read or Lost-Update). To avoid these problems, a transaction manager strives to fulfill the ACID properties, which can be guaranteed by

different approaches. Two-phase locking uses locks to schedule parallel transactions with the risk of deadlocks; an optimistic concurrency control allows all operations to be executed but aborts a transaction if a critical situation is identified; snapshot isolation provides a consistent view (snapshot) to a transaction and checks at commit time whether there is a conflict with another update in the meantime.

Distributed Systems and Transaction Management

NoSQL systems often have relaxed guarantees for transaction management and do not strictly follow the ACID properties. As they focus on availability, their transaction model is abbreviated by BASE: Basically Available, Soft-state, and Eventual consistent. This model is based on the CAP theorem (Brewer 2000), which states that a distributed (database management) system can only guarantee two out of the three properties: consistency, availability, and partition tolerance. To guarantee availability the system has to be robust against failures of nodes (either simple nodes or coordinator nodes). For partition tolerance, the system must be able to compensate network failures. Consistency in distributed DBMS can be viewed from different angles. For a single node, the data should be consistent, but in a distributed system, the data of different nodes might be inconsistent. Eventual consistency, as implemented in many NoSQL systems, assures that the changes are distributed to the nodes, but it is not known when. This provides a better performance as delays because of the synchronization between different nodes is not necessary while the transaction is committed.

In addition to the relaxed transaction management, distributed NoSQL systems provide additional performance by using sharding and replication. Sharding is partitioning the database into several subsets and distributing these shards across several nodes in a cluster. Thus, complex queries processing a huge amount of data (as often required in Big Data applications) can be distributed to multiple nodes, thereby multiplying the compute resources which are available for query processing. Of course, distributed query

processing can be only beneficial for a certain type of queries, but the Map-Reduce programming model fits very well to this type of distributed data management. Replication is also an important aspect for a distributed DBMS which allows for the compensation of network failures. As shards are replicated to several nodes, in the case of failure of a single node, the work can be re-assigned to another node holding the same shard.

DBMS Categories

Relational DBMS

Relational DBMS are by far the most popular, mature, and successful DBMS. Major players in the market are Microsoft SQL Server, Oracle, IBM DB2, MySQL, and PostgreSQL. Their implementations are based on a mathematically well-founded relational data model and the corresponding languages for querying the data (relational algebra and relational calculus) (Codd 1970). A relational database system requires the definition of a schema, before data can be inserted into a database. Main principles in the design of relational schemata are integrity, consistency, and the avoidance of redundant data. The normalization theory has been developed as a strict methodology for the development of relational schemata, which guarantees these principles.

Furthermore, this methodology ensures that the resulting schema is independent of a particular set of queries, i.e., all types of queries or applications can be supported equally well. Relational query languages, mainly SQL and its dialects, are based on the set-oriented relational algebra (procedural) and the relational calculus (declarative). However, one critique of the NoSQL community is that normalized schemata require multiple costly join operations in order to recombine the data that has been split across multiple tables during normalization. Nevertheless, the strong mathematical foundation of the relational query languages allows for manifold query optimization methods on different levels in the DBMS.

NoSQL Database Management Systems

As stated above, NoSQL DBMS have been developed to address several limitations of relational DBMS. The development of NoSQL DBMS started in the early 2000s, with an increasing demand for simple, distributed data management in Internet applications. Limited functions for horizontal scalability and high costs for mature distributed relational DBMS were main drivers of the development of a new class of DBMS. In addition, the data model and query language was considered as too complex for many Internet applications. While SQL is very expressive, the implementation of join queries is costly for an application developer as updates cannot be performed directly on the query result. This leads to the development of new data models that fit better to the need of Internet applications. The main idea of the data models is to store the data in an aggregated data object (e.g., a JSON object or XML document) rather than in several tables as in the relational model.

The simplest data model is the *key-value* data model. Any kind of data object can be stored with a key in the database. Retrieval of the object is done by the key, may be in combination with some kind of path expression to retrieve only a part of the data object. The data object is often represented in the JSON format as this is the data model for web applications. This model is well suited for applications which retrieve objects only by a key, such as session objects or user profiles. On the other hand, more complex queries to filter or aggregate the objects based on their content are not directly supported. The most popular key-value DBMS is Redis (<https://redis.io/>).

The *document-oriented* data model is a natural extension of the key-value model as it stores JSON documents (as a collection of single JSON objects). The query language of document-oriented DBMS is more expressive than for key-value systems, as also filtering, aggregation, restructuring operations can be supported. The most prominent system in this class is MongoDB (<https://www.mongodb.com/>), which also supports a kind of join operation between JSON objects.

Wide column stores can be considered as a combination of relational and key-value DBMS. The logical model of a wide column store has also tables and columns, as in a relational DBMS. However, the physical structure is more like a two-level key-value store. In the first level, a row key is used that links the row key with several column families (i.e., groups of semantically related columns). The second level relates a column family with values for each column. Physically, the data within a column family is stored row-by-row as in relational systems. The advantage is that a column family is only a subset of the complete table; thus, only a small part of the table has to be read if access to only one column family is necessary. Apache Cassandra (<http://cassandra.apache.org/>) is the most popular system in this category.

Finally, *graph-oriented* DBMS store the data in graphs. As described above, graph DBMS can be used in cases where more complex, semantically rich data has to be managed, e.g., data of knowledge graphs. Although, it is often stated that the graph data model fits well to social networking applications, Facebook uses a custom, distributed variant of MySQL to manage their data. Graph-oriented DBMS should be applied only if specific graph analytics is required in the applications (e.g., shortest path queries, neighborhood queries, clustering of the network). Neo4j (<https://neo4j.com/>) is frequently used as graph database.

Streaming

A specific type of DBMS has evolved as data sources producing data at high frequencies became more and prevalent. Low-cost sensors and high-speed (mobile) Internet available to the wide mass of people opened up new possibilities for applications, such as high-speed trading or near real-time prediction. Common DBMS were no longer able to handle nor store these amounts and frequency of data, such the new paradigm of data stream management systems developed (Arasu et al. 2003). A data stream is usually unbounded, and it is not known if and when a stream will end. DSMS are human-passive machine-active systems, i.e., queries are

registered once at the system (hence, also termed standing queries) and are executed over and over again producing results, while data streams into the system. Usually a DSMS follows also a certain data model and usually this is the relational model. Here a stream is equivalent to a table, for which a schema with attributes of a certain domain is defined. Timestamps are crucial for the processing and are an integral part of the schema. Several principles which are valid for common DBMS cannot be applied to DSMS. For example, specific query operators, such as joins, block the production of a result, as it waits infinitely of the stream (the data set) to end to produce a result. Hence, either operator implementations suitable for streams are defined, or only a part of the stream (a window) is used for the query to operate on. Furthermore, there are operators or system components which are responsible to keep up the throughput and performance of the system. If measured QoS parameters indicate that the system gets slower, and if completeness of data is not an issue, tuples may be dropped by sampling them. Parallel to the DSMS paradigm, the concept of complex event processing (CEP) developed. In CEP each tuple is regarded as an event and simple and complex events (describing a situation or a context) are distinguished. CEP systems are specifically designed to detect complex events based on these simple event streams using patterns. CEP systems can be build using DSMS and both share similar or the same technologies to enable stream processing.

Further possibilities to work with streaming data are time series databases (TSDB). TSDB are optimized to work on time series and offer functionality to enable usual operations on time series, such as aggregations over large time periods. In contrast to DSMS, TSDB keep a temporary history of data points for analysis. Here also different data models are used, but more balanced between NoSQL systems and relational systems. They can be operated as in-memory systems or persistent storage systems. Important features are the distributability and clusterability to ensure a high performance and a high availability. TSDBs can furthermore be distinguished based on

the granularity of data storage they offer. A prominent example is InfluxDB included in the TICK stack offering a suite of products for the complete chain for processing, analysis, and alerting. Another example system is Druid, which offers the possibility of OLAP on time series data used by big companies, such as Airbnb or eBay.

Other Data Management Systems in the Context of Big Data

Apache Hadoop (<https://hadoop.apache.org/>) is a set of tools which are focused on the processing of massive amounts of data in a parallel way. The main components of Hadoop are the Map-Reduce programming framework and the Hadoop Distributed File System (HDFS). HDFS, as the name suggests, is basically a file system and can store any kind of data, including simple formats such as CSV and unstructured texts. As a distributed file system, it implements also the features of sharding and replication described above. Thereby, it fits well to the Map-Reduce programming model to support massive parallel, distributed computing tasks.

As HDFS provides the basic functionality of a distributed file system, many NoSQL systems can work well with HDFS as the underlying file system, rather than a common local file system. In addition, specific file formats such as Parquet or ORC have been developed to fit better to the physical structure of data in a HDFS. On top of HDFS, systems like HBase and Hive can be used to provide a query interface to the data in HDFS, which is similar to SQL.

Apache Spark (<https://spark.apache.org/>) is not a DBMS although it also provides data management and query functionalities. In its core, Apache Spark is an analytics engine which can efficiently retrieve data from various backend DBMS, including classical relational systems, NoSQL systems, and HDFS. Spark supports its own dialect of SQL, called SparkSQL, which is translated to the query language of the underlying database system. As Spark has an efficient distributed computing system for transformation or analysis of Big Data, it has become very popular for Big Data applications.

Elasticsearch (<https://www.elastic.co/products/elasticsearch>) is also not a DBMS in the first place, but a search engine. However, it can be used to manage any kind of documents, including JSON documents, thereby, enabling the management of semi-structured data. In combination with other tools to transform (Logstash) and visualize (Kibana), it is often used as a platform for analyzing and visualizing semi-structured data.

We are able to mention here the most important Big Data management systems. As the field of Big Data management is very large and developing very quickly, an enumeration can be only incomplete.

Cross-References

- [Big Data Theory](#)
- [Complex Event Processing \(CEP\)](#)
- [Data Processing](#)
- [Data Storage](#)
- [Graph-Theoretic Computations/Graph Databases](#)
- [NoSQL \(Not Structured Query Language\)](#)
- [Semi-structured Data](#)
- [Spatial Data](#)

References

- Abiteboul, S., Buneman, P., & Suciu, D. (1999). *Data on the web: From relations to semistructured data and XML*. San Francisco: Morgan Kaufmann.
- ACID Transaction. (2009). In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 21–26). Springer US. Retrieved from https://doi.org/10.1007/978-0-387-39940-9_2006.
- Arasu, A., Babu, S., & Widom, J. (2003). An abstract semantics and concrete language for continuous queries over streams and relations. In Proceedings of international conference on data base programming languages.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 34–43.
- Brewer, E. A. (2000). Towards robust distributed systems (abstract). In G. Neiger (Ed.), Proceedings of the nineteenth annual ACM symposium on principles of distributed computing, 16–19 July 2000, Portland. ACM, p. 7.
- Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377–387.

- Elmasri, R. A., & Navathe, S. B. (2017). *Fundamentals of database systems* (7th ed.). Harlow: Pearson.
- Härder, T. (2005). DBMS architecture – still an open problem. In G. Vossen, F. Leymann, P. C. Lockemann, & W. Stucky (Eds.), *Datenbanksysteme in business, technologie und web, 11. fachtagung des gi-fachbereichs “datenbanken und informationssysteme” (dbis), karlsruhe, 2.-4. märz 2005* (Vol. 65, pp. 2–28). GI. Retrieved from <http://subs.emis.de/LNI/Proceedings/Proceedings65/article3661.html>.
- Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. San Rafael: Morgan & Claypool Publishers.

Datacenter

► [Data Center](#)

Data-Driven Discovery

► [Data Discovery](#)

Datafication

Clare Southerton
Centre for Social Research in Health and Social
Policy Research Centre, UNSW, Sydney, Sydney,
NSW, Australia

Definition

Datafication refers to the process by which subjects, objects, and practices are transformed into digital data. Associated with the rise of digital technologies, digitization, and big data, many scholars argue datafication is intensifying as more dimensions of social life play out in digital spaces. Datafication renders a diverse range of information as machine-readable, quantifiable data for the purpose of aggregation and analysis. Datafication is also used as a term to describe a logic that sees

things in the world as sources of data to be “mined” for correlations or sold, and from which insights can be gained about human behavior and social issues. This term is often employed by scholars seeking to critique such logics and processes.

Overview

The concept of datafication was initially employed by scholars seeking to examine how the digital world is changing with the rise of big data and data economies. However, as datafication itself becomes more widespread, scholarship in a range of disciplines and subdisciplines has drawn on the concept to understand broader shifts towards rendering information as data for pattern analysis, beyond online platforms. The concept of datafication, in its present form, emerged in the last 5 years with the growth of data analytics, being popularized by Viktor Mayer-Schönberger and Kenneth Cukier’s (2013) book *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, who describe its capacity as a new approach for social research. While datafication is distinct from digitization, as Mayer-Schönberger and Cukier point out (2013), digitization is often part of the process of datafication. So too is quantification, as when information is “datafied,” it is reduced to elements of the information that can be counted, aggregated, calculated, and rendered machine-readable. As such there are significant complexities that are lost in this process that renders qualitative detail invisible, and indeed this critique has been substantially developed in the literature examining the logics of big data (see, e.g., Boyd and Crawford 2011; Kitchin 2014; van Dijck 2014). Furthermore, a range of methodological and epistemological issues are raised about the insights of data drawn from new data economies, in which there are a range of existing inequalities, as well as huge value to be found in encouraging participation in digitally mediated social interaction and practices of sharing personal information online (see, e.g., Birchall 2017; van Dijck 2013; Zuboff 2015).

The Datalogical Turn

As more and more aspects of social life have begun to generate digital data, the possibility of analyzing this data in order to produce opportunities for profit has substantially changed the nature of how digital infrastructures are oriented. In particular, the capacity that exists now to analyze large data sets, what we call ‘big data’, and the ability to draw together insights from multiple data sets (e.g., search engine data, social media demographic information, viewing history on YouTube etc.) has significantly changed how online platforms operate. Data scientists seek to produce findings on a wide range of issues by examining the data traces that individuals left behind. The big data analytics have been used in the private sector for a range of purposes including for filtering digital content or products in the online marketplace in the form of recommendations, and, most prominently, through targeted advertisements. In addition, datafication has been identified by some as an opportunity to gain unprecedented access to data for social research. Sometimes called “the datalogical turn” or “the computational turn,” recently greater attention has been paid to the sociological insights offered by these large datasets. There has also been unease in the social sciences surrounding the use of big data, particularly social media data, to analyze culture and address social problems. Media scholar José van Dijck (2014) argues that datafication has become a pervasive ideology – “dataism” – in which the value and insights of aggregated data are seen as implicit. This ideology also places significant trust and legitimacy in the institutions that collect this data, despite their financial interests. Indeed, such an ideology is clear in the claims made early in the big data revolution about the exponential capacity of big data to explain social life, with some proponents of big data analysis proclaiming the “end” of social theory. These claims were made on the basis that theorizing *why* people act in certain ways, indeed the very questions that formed the basis of much social scientific inquiry, was rendered irrelevant by big data’s ability to see patterns of actions on a mass scale. In essence,

datafication was seen as a way to bypass the unnecessary complexity of social life and identify correlations, without the need for meaningful explanation. Many of these early claims have been tempered, especially as the predictive power of big data has failed to deliver on many of its utopian promises.

The Datafication of Social Life

The data that is aggregated by data scientists is predominantly made possible by the data traces that online interactions generate which can now be collected and analyzed, by what we might term the “datafication of social life.” As social media platforms have come to host more of our daily interactions, these interactions have become parcels of data in the form of comments, likes, shares, and clicks. Alongside these interactions, our digital lives are also comprised of a range of data-gathering activities: web browsing and using search engines, interactions with advertisements, online shopping, digital content streaming, and a vast array of other digital practices are rendered as pieces of data that can be collated for analysis to identify trends and, likely if commercialized, further opportunities for profit (Lupton 2019). Even beyond the activities users undertake online, the geo-locative capacities of digital devices now allow the collection of detailed location data about their user. This datafication of social life has significantly changed the organization of digital platforms as profit can now be drawn by the collection of data, and as such dataveillance has become embedded into almost all aspects of digital interaction.

Beyond online interactions and social media use, recent years have seen datafication and data analytics spread to a range of fields. Scholars have identified the datafication of health, both in the trend of individual self-tracking technologies, such as fitness trackers and smart watches, and the ways in which clinical practice has become increasingly data-driven, especially when it comes to how governments deal with medical information (Ruckenstein and Schüll 2017). So too has education been impacted by datafication,

as children are increasingly monitored in schools by RFID in uniforms, facial recognition-enabled CCTV, and online monitoring of classwork (Taylor 2013). Scholars have also drawn attention to the forms of dataveillance impacting childhood beyond education, through parenting apps and child-tracking technologies. The spread of datafication points to the power of the pervasive ideology of datafication van Dijck (2014) described, whereby objective truth is to be found by rendering any social problem as digital data for computational analysis.

Critiques of Datafication

The logics of datafication have been substantially critiqued by social scientists. Privacy and surveillance scholars have highlighted widespread issues surrounding the way datafication facilitates the collection of personal information passively, in ways that platform users may not be aware of, and data is stored for a wide range of future uses that users cannot meaningfully consent to. As datafication spreads into more areas of social life, notions of consent become less helpful as users of digital platforms and datafied services often feel they do not have the option to opt out. Furthermore, large-scale data leaks and hacks of social media platforms demonstrate the fragility of even high-standard data protection systems.

In addition to privacy concerns, datafication can reproduce and even exacerbate existing social inequalities. Data-driven risk evaluation systems such as those now routinely employed by financial service providers and insurance companies can perpetuate discrimination against already marginalized communities (Leurs and Shepherd 2017). Furthermore, such discrimination is masked by the mythology of objectivity, insight, and accuracy surrounding these systems, despite their often opaque workings. While discrimination is certainly not new, nor does it arise solely as a product of datafication and systems drive by big data, however these systems facilitate discrimination in a manner that eludes observation and dangerously legitimizes inequalities as natural, evidenced by data, rather than a product of implicit bias.

Scholars have also raised concerns about the datafication of social science and the ways computational methods have impacted social research. Computational social science has been accused of presenting big data as speaking for itself, as a kind of capture of social relations rather than constituted by commercial forces and indeed by the new forms of digital sociality this data emerges from. For example, using social media data as a way to gauge public opinion often inappropriately represents such data as representative of society as a whole, neglecting important differences between the demographics of different platforms and the specific affordances of digital spaces, which may give rise to different responses. Similarly, these large data can establish correlations between seemingly disparate variables that, when presented as proof of a causal relationship, can prove misleading (Kitchin 2014). This is not to suggest scholars disregard this data, but rather caution must be employed to ensure that such data is appropriately contextualized with theoretically informed social scientific analysis. Qualitative differences must be examined through attention rather than smoothed out.

Conclusion

The process of datafication serves to transform a wide range of phenomena into digitized, quantifiable units of information for analysis. With the mass infiltration of smart technologies into everyday life and as more social interaction is filtered through social media platforms and other online services, data is now generated and collected from a diverse array of practices. Consequently, datafication and computational social science can offer significant insights into digitally embedded lives. However, as many scholars in the social sciences have argued, inevitably this process is reductive of the complexity of the original object and the rich social context to which it belongs.

Further Reading

Birchall, C. (2017). *Shareveillance: The dangers of openly sharing and covertly collecting data*. Minneapolis: University of Minnesota Press.

- boyd, d., & Crawford, K. (2011). *Six Provocations for Big Data*. Presented at the a decade in internet Time: Symposium on the Dynamics of the Internet and Society, Oxford. <https://doi.org/10.2139/ssrn.1926431>.
- Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12.
- Leurs, K., & Shepherd, T. (2017). Datafication & Discrimination. In M. T. Schäfer & K. van Es (Eds.), *The Datafied society: Studying culture through data* (pp. 211–234). Amsterdam: Amsterdam Press.
- Lupton, D. (2019). *Data selves: More-than-human perspectives*. Cambridge: Polity.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. London: John Murray Publishers.
- Ruckenstein, M., & Schüll, N. D. (2017). The Datafication of health. *Annual Review of Anthropology*, 46(1), 261–278.
- Taylor, E. (2013). Surveillance schools: A new era in education. In E. Taylor (Ed.), *Surveillance schools: Security, discipline and control in contemporary education* (pp. 15–39). London: Palgrave Macmillan UK.
- van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford: Oxford University Press.
- van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197–208.
- Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology Impact*, 30(1), 75–89.

Data-Information-Knowledge-Action Model

Xiaogang Ma

Department of Computer Science, University of Idaho, Moscow, ID, USA

Synonyms

DIKW pyramid; Information hierarchy; Knowledge hierarchy; Knowledge pyramid

Introduction

Facing the massive amounts and various subjects of datasets in the Big Data era, it is impossible for humans to handle the datasets alone. Machines are needed in data manipulation, and a model of Data-Information-Knowledge-Action will help guide

us through the process of applying big data to tackle scientific and societal issues. Knowledge is one's expertise or familiarity with a subject under working. Knowledge is necessary in the process to generate information from data about a certain issue, and then take actions. New knowledge can be generated on both the individual level and the community level, and certain explicit knowledge can be encoded as machine readable knowledge bases and be used as tools to facilitate the process of data management and analysis.

Understand the Concepts

The four concepts data, information, knowledge and action are often seen in the language people used in problem tackling and decision-makings for various scientific and societal issues. Data are the representation of some facts. We can see data of various topics, types, and dimensions in the real world, such as a geologic map of United Kingdom, records of sulfur dioxide concentration in the plume of Poás Volcano, Costa Rica, the weekly records of the sales of cereal in a Wal-Mart store located at Albany, NY, and all the Twitter tweets with hash tag #storm in January 2015. Data can be recorded on different media. The computer hard disks, thumb drives, and CD-ROMS that are popularly used in nowadays are just part of the media, and the use of computer readable media significantly promote and speed-up the transmission of data.

Information is the meaning of data as interpreted by human beings. For example, a geologist may find some initial clues for iron mine exploration by using a geologic map, a volcanologist may detect a few abnormal sulfur dioxide concentration values about the plume of a volcano, a business manager may find that the sales of the cereal of a certain brand have been rising in the past three weeks, and a social media analyst may find spatio-temporal correlations between tweets with hash tag #storm and the actual storm that happened in northeast United States in January, 2015. In the context of Big Data, there are massive amounts of data available but most of them could be just noise and are irrelevant to the subject under working. In this situation, a step of

data cleansing can be deployed to validate the records, remove errors, and even collect new records to enrich the data. People then need to discover clues, news, and updates that make sense to the subject from the data.

Knowledge is people's expertise or familiarity with one or more subjects under working. People use their knowledge to discover information from data, and then make decisions based on the information and their knowledge. The knowledge of a subject is multifaceted and is often evolving with new discoveries. The process from data to information in turn may make new contribution to people's knowledge. A large part of those expertise and familiarity are often described as tacit knowledge in a people's brain, which is hard to be written down or shared. In contrast, it is now possible to record and encode a part of people's knowledge into machine readable format, which is called explicit or formal knowledge, such as the controlled vocabulary of a subject domain. Such recorded formal knowledge can be organized as a knowledge base, which, in conjunction with an inference engine, can be used to build an expert system to deduce new facts or detect inconsistencies.

Action is the deeds or decisions made based on knowledge and information. In real-world practices of the Data-Information-Knowledge-Action model, data are collected surrounding a problem to be addressed, then the data are interpreted to identify competing explanations for the problem, as well as uncertainties of the explanations. Based on those works, one or more decisions can be made, and the decisions will be reflected in following actions. The phrase "informed decision" is often used nowadays to describe such a situation that people make decisions and take actions after knowing all the pros and cons and uncertainties based on their knowledge as well as the data and information in hand.

Mange the Relationships

In practice, data users such as scientific researchers and business managers need to take

care of the relationships among the four concepts in the Data-Information-Knowledge-Action model as well as the loops between two or more steps in the usage of this model. The first is the data-information relationship. The saying "Garbage in, garbage out" describes the relationship in an intuitive way. Users need to recognize the inherent flaws of all data and information. For example, the administrative boundary areas on geological maps often confuse people because the geological survey and mapping of the areas at each side of a boundary are taken by different teams, and there could be discrepancies in their description about the same region that is divided by the boundary. For complex datasets, visualization is a useful approach, which can be used not only for summarizing data to be processed but also for presenting the information generated from the data (Ma et al. 2012).

The second relationship is between information and knowledge. People may interpret the same data in different ways based on their knowledge, and generate different information. Even for a same piece of information, they may use it differently in decision-making. To use data and information in a better way, it is necessary for people to keep in mind about what they know best and what they know less, and collaborate with colleagues where necessary. If information is generated from data through a workflow, well-documented metadata about the workflow will provide provenance for both data and information, which improve the credibility of them. Through reading the metadata, users know about the data collectors, collection date, instruments used, data format, subject domains, as well as the experiment operator, software program used, methods used, etc., and thus obtain insight into the reliability, validity, and utility of the data and information.

The third relationship is between knowledge and action. For a real-world workflow, the focus to address the facing problems is the action items. Analyzing and solving problems need knowledge on many subjects. Knowledge sharing is necessary within a working team that runs the workflow. Team members need to come together

to discuss essential issues and focus on the major goals. For example, a mineral exploration project may need background knowledge in geology, geophysics, geochemistry, remote sensing image processing, etc. because data of those subjects are collected and to be used. People need to contribute their knowledge and skills relevant to the problem and collaborate with others to take actions. In certain circumstances, they need to speed up the decision and action process at the expense of optimization and accuracy.

An ideal usage of the Data-Information-Knowledge-Action model is that relevant information is discovered, appropriate actions are taken, and people's knowledge is enhanced. However, in actual works there could be closed loops that weaken the usefulness of the model. For example, there could be loops among data, information and knowledge without actions taken. There could also be loops between knowledge and action, in which no information from latest data is used in decision-making. Another kind of loops is between information and action, in which no experience is saved to enhance the knowledge and people continue to respond to the latest information without learning from previous works.

Reverse Thinking

Besides relationships and loops between two or more steps in the usage of the Data-Information-Knowledge-Action model, the model itself can also be considered in a reversed way. The so-called knowledge-based system or expert system is the implementation of the reverse thinking of this model. There are two typical components in a knowledge-based system, one is a knowledge base that encodes explicit knowledge of relevant subject domains and the other is an inference engine. While people's knowledge is always necessary in the process to discover information from data, the knowledge-based systems are a powerful tool to facilitate the process. In the field of semantic web, people are working on ontologies as a kind of knowledge base. An

ontology is a specification of a shared conceptualization of a domain. There are various types of ontologies based on the level of details on the specification of the meaning of concepts and the assertion of relationships among concepts. For instance, a controlled vocabulary of a subject domain can be regarded as a simple ontology. Ontologies are widely used in knowledge-based systems to handle datasets in the Web, including the generation of new information. In a recent research, an ontology was built for the geologic time scale. Geologic time is reflected in the rock age descriptions on various geologic map services. The built ontology was used to retrieve the spatial features of with records of certain rock ages and to generalize the spatial features according to user commands. To realize the map generalization function, programs were developed to query the relationships among geologic time concepts in the ontology. Besides those functions, the ontology was also used to annotate rock age concepts retrieved from a geologic map service based on the concept specifications encoded in the ontology and further information retrieved from external resources such as the Wikipedia.

Another kind of reverse thinking is that, after a whole procedure of Data-Information-Knowledge-Action, changes in knowledge may take place inside an individual who took part in the workflow. The changes could be the discovery of new concepts, recognition of updated relationships between concepts, or modification of previous beliefs, etc. The word "action learning" is used to describe the situation that an individual learns when he takes part in an activity. When the individual has learned to do a different action, he has obtained new knowledge. This can be regarded as a reversed step to the Data-Information-Knowledge-Action model and can also be regarded as an extension to it. That is, two other steps, Learning and New Knowledge can be added next to the Action step in the model. All learning is context dependent (Jensen 2005). Similar to the aforementioned collaboration among individuals to transform data into information, learning takes place as a negotiation of

meaning between the collaborators in a community of practice. Individuals learn and obtain new knowledge, and in turn the new knowledge can be used to discover new information from data, which provides the foundation for the creation of new knowledge in other individuals' minds.

Communities of Practice

The generation of new knowledge involves human thinking with information. The significant progress of artificial intelligence and knowledge-based systems in recent years has inspired knowledge revolution in various subject domains (Petrides 2002). Yet, the knowledge revolution itself still needs human systems to realize it. To facilitate knowledge revolution, issues relevant to thinking and information should both be addressed. An intuitive and natural way to do this is to build communities and promote communities of practice. As mentioned above, communities of practice not only allow members to work together on data analysis to generate new information, they also help community members think together to generate new knowledge, on both the individual level and the community level (Clampitt 2012). Modern information technologies have already provided efficient facilities for such collaborations, and more challenges are from the social or cultural side. That is, individuals in a community should be willing to share their findings and be open to new ideas. For the community as whole, it should maintain diversity while trying to achieve consensus on the commonly shared and agreed ideas (McDermott 1999). A typical example for such communities is the World Wide Web Consortium (W3C), which develops standards for the Web. A large part of W3C's work is coordinating the development of ontologies for various domains, which can be regarded as machine readable knowledge bases. An ontology is normally under work by a group of individual and organizations across the world and should go through several stages of review, test and revision before it can become a W3C Recommendation. The construction and implementation of ontologies

significantly improve the interoperability of multi-source datasets made open on the web, and facilitate the development of intelligent functions to aid the procedure of Data-Information-Knowledge-Action in Big Data manipulation.

Cross-References

- [Data Provenance](#)
- [Data-Information-Knowledge-Wisdom \(DIKW\) Pyramid, Framework, Continuum](#)
- [Decision Theory](#)
- [Knowledge Management](#)
- [Pattern Recognition](#)

Further Reading

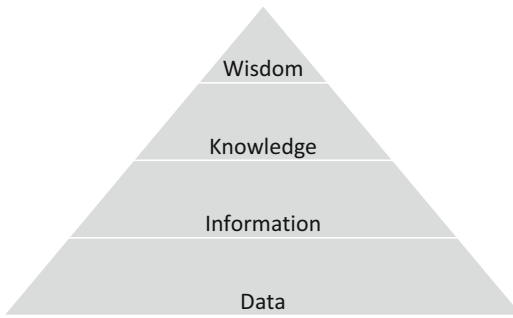
- Clampitt, P. G. (2012). *Communicating for managerial effectiveness: Problems, strategies, solutions*. Thousand Oaks: SAGE Publications.
- Jensen, P. E. (2005). A contextual theory of learning and the learning organization. *Knowledge and Process Management*, 12(1), 53–64.
- Ma, X., Carranza, E. J. M., Wu, C., & van der Meer, F. D. (2012). Ontology-aided annotation, visualization and generalization of geological time scale information from online geological map services. *Computers & Geosciences*, 40, 107–119.
- McDermott, R. (1999). Why information technology inspired but cannot deliver knowledge management. *California Management Review*, 41(4), 103–117.
- Petrides, L. A. (2002). Organizational learning and the case for knowledge-based systems. *New Directions for Institutional Research*, 2002(113), 69–84.

Data-Information-Knowledge-Wisdom (DIKW) Pyramid, Framework, Continuum

Martin H. Frické

University of Arizona, Tucson, AZ, USA

The Data-Information-Knowledge-Wisdom (DIKW) hierarchy, or pyramid, relates data, information, knowledge, and wisdom as four layers in a



Data-Information-Knowledge-Wisdom (DIKW) Pyramid, Framework, Continuum, Fig. 1 The knowledge pyramid

pyramid. Data is the foundation of the pyramid, information is the next layer, then knowledge, and, finally, wisdom is the apex. DIKW is a model or construct that has been used widely within Information Science and Knowledge Management. Some theoreticians in library and information science have used DIKW to offer an account of logico-conceptual constructions of interest to them, particularly concepts relating to knowledge and epistemology. In a separate realm, managers of information in business process settings have seen the DIKW model as having a role in the task meeting real world practical challenges involving information (Fig. 1).

Historically, the strands leading to DIKW come from a mention by the poet T.S. Eliot and, separately, from research from Harland Cleveland and the systems theorists Mortimer Adler, Russell Ackoff, and Milan Zeleny. The main views are perhaps best expressed in the traditional sources of Adler, Ackoff, and Zeleny. Russell Ackoff, in his seminal paper, describes the pyramid from the top down:

Wisdom is located at the top of a hierarchy of types . . . Descending from wisdom there are understanding, knowledge, information, and, at the bottom, data. Each of these includes the categories that fall below it . . . (Ackoff 1989, 3)

In fact, the way the pyramid works as a method is from the bottom up, not top down. The process starts with data and ascends to wisdom.

Data are the symbolic representations of observable properties

Data are symbols that represent properties of objects, events and their environments. They are products of observation. To observe is to sense. The technology of sensing, instrumentation, is, of course, highly developed. (Ackoff 1989, 3)

In turn, information is relevant, or usable, or significant, or meaningful, or processed, data (Rowley 2007, Section 5.3 Defining Information). The vision is that of a human asking a question beginning with, perhaps, “who,” “what,” “where,” “when,” or “how many” (Ackoff 1989, 3); and the data is processed into an answer to an enquiry. When this happens, the data becomes “information.” Data itself is of no value until it is transformed into a relevant form.

Information can also be inferred from data – it does not have to be immediately available. For example, were an enquiry to be “what is the average temperature for July?”; there may be individual daily temperatures explicitly recorded as data, but perhaps not the average temperature; however, the average temperature can be calculated or inferred from the data about individual temperatures. The processing of data to produce information often reduces that data (because, typically, only some of the data is relevant). Ackoff writes

Information systems generate, store, retrieve, and process data. In many cases their processing is statistical or arithmetical. In either case, information is inferred from data. (Ackoff 1989, 3)

Information is relevant data, together with, on occasions, the results of inferences from that relevant data. Information is thus a subset of the data, or a subset of the data augmented by additional items inferred or calculated or refined from that subset.

Knowledge, in the setting of DIKW, is often construed as know-how or skill. Ackoff suggests that know-how allows an agent to promote information to a controlling role – to transform information into instructions.

Knowledge is know-how, for example, how a system works. It is what makes possible the transformation of information into *instructions*. It makes *control* of a system possible. To control a system is to make it work *efficiently*. (Ackoff 1989, 4)

Next up the hierarchy are understanding and wisdom. The concept of understanding is almost always omitted from DIKW by everyone (except Ackoff) and, in turn, wisdom is given only limited discussion by researchers and theorists. Ackoff sees wisdom as being the point at which humans inject ethics or morality into systems. He explains

Wisdom adds value, which requires the mental function we call judgement.... The value of an act is never independent of the actor... [ethical and aesthetic values] are unique and personal.

...wisdom-generating systems are ones that man will never be able to assign to automata. It may well be that wisdom, which is essential to the effective pursuit of ideals, and the pursuit of ideals itself, are the characteristics that differentiate man from machines. (Ackoff 1989, 9)

Ackoff concludes with some numbers, apparently produced out of thin air without any evidence

...on average about forty percent of the human mind consists of data, thirty percent information, twenty percent knowledge, ten percent understanding, and virtually no wisdom. (Ackoff 1989, 3)

Modern Developments and Variations

There are publications that argue that DIKW should be top down in process, not bottom up. The suggestion is that there is no such thing as “raw data,” rather all data must have theory in it and thus theory (i.e., knowledge, information) must illuminate data, top down, rather than the other way around, bottom up. There are publications that add or subtract layers from DIKW—most omit understanding, some omit wisdom, some add messages and learning, and there are other variations on the theme. There are publications that draw DIKW more into management practices, into business process theory, and into organizational learning. Finally, there are publications that take DIKW into other cultures, such as the Maori culture.

Drawing It All Together and Appraising DIKW

DIKW seems not to work. Ackoff (1989) urges us to gather data with measuring instruments and sensors. But instruments are constructed in the light of theories, and theories are essential to inform us of what the surface indications of the instruments are telling us about a reality beyond the instruments themselves. Data is “theory-laden.” Data itself can be more than the mere “observable,” and it can be more than the pronouncements of “instruments.” There are contexts, conventions, and pragmatics at work. In particular circumstances, researchers might regard some recordings as data which report matters that are neither observable nor determinable by instrument.

All data is information. However, there is information that is not data. Information can range much more widely than data; it can be much more extensive than the given. For example, consider the universal statements “All rattlesnakes are dangerous” or “Most rattlesnakes are dangerous.” These statements presumably are, or might be, information, yet they cannot be inferred from data. The problem is with the universality, with the “All” or “More.” Any data, or conjunctions of data, are singular. For example, “Rattlesnake A is dangerous,” “Rattlesnake B is dangerous,” “Rattlesnake C is dangerous,” etc., are singular in form. Trying to make the inference from “some” to “all,” or to “most,” are inductive inferences, and inductive inferences are invalid.

The step from information to knowledge is also not the easiest. In epistemology, philosophers distinguish knowledge that from knowledge how. A person might know *that* the Eiffel Tower is in France, and she might also know *how* to ride a bicycle. If knowledge is construed as “know-that,” then, under some views of information and knowledge, information and knowledge are much the same. In which case, moving from information to knowledge might not be so hard. However, in the context of DIKW, knowledge is usually taken to be “know-how,” and that makes the step difficult. Consider a young person learning how to ride

a bike. What information in particular is required? It is hard to say, and maybe no specific information in particular is required. However, like many skills, riding a bike is definitely coachable, and information can improve performance. Know-how can benefit from information. The problem is in the details.

Wisdom is in an entirely different category to data, information, and know-how. Wisdom certainly uses or needs data, information, and know-how, but it uses and needs more besides. Wisdom is not a distillation of data, information, and know-how. Wisdom does not belong at the top of a DIKW pyramid. Basically, this is acknowledged implicitly by all writers on the topic, from Plato, through Ackoff, to modern researchers.

What about DIKW in the setting of work processes? The DIKW theory seems to encourage uninspired methodology. The DIKW view is that data, existing data that has been collected, is promoted to information and that information answers questions. This encourages the mindless and meaningless collection of data in the hope that one day it will ascend to information – i.e., preemptive acquisition. It also leads to the desire for “data warehouses,” with contents that are to be analyzed by “data mining.” Collecting data also is very much in harmony with the modern “big data” approach to solving problems. Big data and data mining are somewhat controversial. The worry is that collecting data *blind* is suspect methodologically.

Know-how in management is simply more involved than DIKW depicts it. As Weinberger (2010) writes

... knowledge is not a result merely of filtering or algorithms. It results from a far more complex process that is social, goal-driven, contextual, and culturally-bound. We get to knowledge — especially “actionable” knowledge — by having desires and curiosity, through plotting and play, by being wrong more often than right, by talking with others and forming social bonds, by applying methods and then backing away from them, by calculation and serendipity, by rationality and intuition, by institutional processes and social roles. Most important in this regard, where the decisions are tough and knowledge is hard to come by, knowledge is not determined by information, for it is the knowing

process that first decides which information is relevant, and how it is to be used.

Wisdom is important in management and decision-making, and there is a literature on this. But, seemingly, no one wants to relate wisdom in management to the DIKW pyramid. The literature on wisdom in management is largely independent of DIKW.

In sum, DIKW does not sit well in modern business process theory. To quote (Weinberger 2010) again

The real problem with the DIKW pyramid is that it's a pyramid. The image that knowledge (much less wisdom) results from applying finer-grained filters at each level, paints the wrong picture. That view is natural to the Information Age which has been all about filtering noise, reducing the flow to what is clean, clear and manageable. Knowledge is more creative, messier, harder won, and far more discontinuous.

Further Reading

- Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16, 3–9.
- KM4DEV. (2012). DIKeW model. Retrieved from http://wiki.km4dev.org/DIKW_model.
- Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*, 33(2), 163–180.
- Weinberger, D. (2010). The problem with the data-information-knowledge-wisdom hierarchy. *Harvard Business Review*. Retrieved from <https://hbr.org/2010/02/data-is-to-info-as-info-is-not>.
- Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4), 479–493. <https://doi.org/10.1002/asi.20508>.

Dataviz

► Data Visualization

Dataviz

► Data Visualization

Decision Theory

Magdalena Bielenia-Grajewska

Division of Maritime Economy, Department of
Maritime Transport and Seaborne Trade,
University of Gdansk, Gdansk, Poland
Intercultural Communication and
Neurolinguistics Laboratory, Department of
Translation Studies, University of Gdansk,
Gdansk, Poland

Decision Theory can be defined as the set of approaches used in various disciplines, such as economics, psychology and statistics, directed at the notions and processes connected with making decisions. Decision theory focuses on how people decide, what determines their choices and what the results of their decisions are. Taking into account the fact that decisions are connected with various spheres of one's life, they are the topic of investigation among researchers representing different disciplines. The popularity of decision theory is not only related to its multi-disciplinary nature but also to the features of the twenty-first century. Modern times are characterized by multitasking as well as the diversity of products and services that demand making decisions more often than before. In addition, the appearance and growth of big data have led to the intensified research on decision processes themselves.

Decision Theory and Disciplines

Decision theory is the topic of interest taken up by the representatives of various disciplines. The main questions interesting for linguists focus on how linguistic choices are made and what the consequences of linguistic selections are. The decisions can be studied by taking different dimensions into account. Analyzing the micro perspective, researchers investigate how words or phrases are selected. This sphere of observation includes the studies on the role of literal and nonliteral language in communication. As Bielenia-Grajewska (2014, 2016) highlights,

research may include, e.g., the role of figurative language in making decisions. Thus, it may be studied how metaphors facilitate or hinder the process of choosing among alternatives. Another scope of investigation is the link between metaphors and emotions in the process of decision-making. Staying within the field of linguistics, education also benefits from decision theories. Understanding how people select among offered alternatives facilitates the cognition of how foreign languages are studied. Decision theory in, e.g., foreign language learning should focus on both learners and teachers. Taking the teacher perspective into account, decision is related to how methods and course materials for students are selected. Looking at the same notion from the learner point of view, investigation concentrates on the way students select language courses and how they learn. Apart from education, linguistic decisions can be observed within organizational settings. For example, managers decide to opt for a linguistic policy that will facilitate effective corporate communication. It may include, among others, the decisions made on the corporate lingo and the usage of regional dialects in corporate settings. Decisions are also made in the sphere of education. Taking into account the studied domain of linguistics, individuals decide which foreign languages they want to study and whether they are interested in languages for general purposes or languages for specific purposes. Another domain relying on decision theory is economics. Discussing economic sub-domains, behavioral economics concentrates on how various determinants shape decisions and the market as such. Both qualitative and quantitative methods are used to elicit needed data. Focusing more on neuroscientific dimensions, neuroeconomics focuses on how economic decisions are made by observing the brain and the nervous system. Management also involves the studies on decision making. The way managers make their decisions is investigated by taking into account, among others, their leadership styles, their personal features, the type of industry they work in and the type of situations they have to face. For example, researchers are interested how managers decide in standard and crisis situations, and how their communication with employees

determines the way they decide. Decision theory is also grounded in politics since it influences not only the options available but also the type of decisions that can be made. For example, the type of government, such as democracy, autocracy and dictatorship determine the influence of individuals on decision-making processes taking place in a given country. As Bielenia-Grajewska (2013, 2015) stresses in her publications, one of other important fields for studying decisions is neuroscience. It should be underlined that the recent advancement in scientific tools has led to the better understanding of how decisions are made. First of all, neuroscience offers a complex and detailed method of studying one's choices. Secondly, neuroscientific tools enable the study of cognitive aspects that cannot be researched by using other forms of investigation. A similar domain focusing on decision theory is cognitive science, with its interest in how people perceive, store information and make decisions. Decision theory is a topic of investigation in psychology; psychologists study how decisions are made or what makes decision-making processes difficult. One of the phenomena studied is procrastination, being the inability to plan decision-making processes, concentrating on less important tasks and often being late with submitting the work on time. Psychologists also focus on heuristics, studying how people find solutions to their problems by trying to reduce the cognitive load connected with decision processes and, consequently, supporting their choices with, e.g., guesses, stereotyping, or common sense. Psychologists, psychiatrists, and neurologists are also interested how decisions are made by people suffering from mental illnesses or having experienced a brain injury. Another domain connected with decision theory is ethics. The ethical aspect of decision-making processes focuses on a number of notions. One of them is providing correct information on offered possibilities. The ethical perspective is also connected with the research itself. It involves not only the proper handling of data gathered in experiments but also using tools and methods that are not invasive for the subjects. If the methods used in experiments carry any type of risk for the participant, the person should be informed about such issues.

Factors Determining Decisions

Another factor shaping decisions is motivation often classified in science by taking the division into intrinsic and extrinsic features into account. For example, motivational factors may be related to one's personal needs as well as the expectation to adjust to the social environment. Another factor determining decisions is technology; its role can be investigated in various ways. One of them is technology as the driver of advancements; thus, technologically advanced products may be purchased more than other merchandise since they limit the amount of time spent on some activities. In addition, technology may influence decision-making processes indirectly. For example, commercials shown on TV stimulate purchasing behaviors. Another application of technology is the possibility of making decisions online. The next factor determining decision-related processes is language. Language in that case can be understood in different ways. One approach is to treat language as the tool of communication used by a given group of people. It may be a national language, a dialect or a professional sublanguage. Taking into account the issue of a language understood in a broad way, being proficient in the language the piece of information is written in facilitates decision processes. On the other hand, the lack of linguistic skills in a given language may lead to the inability of making decisions or making the wrong ones. Apart from the macro level of linguistic issues, the notion of professional sublanguage or specialized communication determines the process of decision-making. One of the most visible examples of the link between professional discourses can be studied in the organizational settings. Taking into account the specialized terminology used in professional discourse, the incomprehension or misunderstanding of specialized discourses determines decision processes. Decisions vary by taking into account the source of knowledge. Decisions can be made by using one's own knowledge or gained expertise through the processes of social interactions, schooling or professional training. In addition, decisions can be divided into individual decisions, made by one person, and cooperative decisions, made by groups of individuals.

Decision may also be forced or voluntary, depending on the issue of free will in decision-making processes. There are certain concepts that are examined through the prism of decision-making. One of them is rationality, studied by analyzing the concept of reason and its influence on making choices. Thus, decisions can be classified as rational or irrational, depending whether they are in accordance with reason, that is whether they reflect the conditions of the reality. Decisions are also studied through the prism of compromise. Researchers take into account the complexity of issues to be decided upon and the steps taken towards reaching the compromise between offered choices. Decision theory also concentrates on the influence of other impulses on decision-making processes. This notion is studied in, e.g., marketing to investigate how auditory (songs or jingles) or olfactory (smells) stimuli determine purchasing behaviors. Apart from these dimensions, marketing specialists are interested in the role of verbal and nonverbal factors in making the selection of merchandise. Decision theory also focuses on the understanding of priorities. As has been mentioned before, psychologists are interested why some individuals procrastinate. Decision theory is focused on different stages of decision-making processes; decisions do not only have to be made, by selecting the best (in one's opinion) alternative, but they should be later implemented and their results should be monitored.

Decision Theory and Methodology

Methodologies used in decision theories can be classified by taking into account different approaches. One of the ways decision theories can be studied is investigating various disciplines and their influence on understanding how decisions are made. As far as the decisions are made, the approaches discussed in different theories turn out to be useful. Thus, such domains as, among others, psychology, economics, and management use theories and approaches that facilitate the understanding of decision-making. For example, psychology relies on cognitive models and behavioral studies. Taking into account linguistics, Critical Discourse Analysis facilitates the creating and

understanding of texts since this type of analysis offers a deep understanding of verbal and nonverbal tools that facilitate decision-making. In addition, intercultural communication, together with its classifications of national or professional cultures, is important in understanding how decisions are made by studying differences across cultures and occupations. Taking into account cross-disciplinary approaches, network and systemic theories offer the understanding of complexities determining decision-making processes. For example, Actor-Network-Theory stresses the role of living and non-living entities in decisional activities. ANT draws one's attention to the fact that also technological advancements, such as the Internet or social media tools influence decisions, by, e.g., offering the possibility of making choices also online. Social Network Analysis (SNA) focuses on the role of relations in individual and group decisions. Thus, analyzing the studied multifactorial and multidisciplinary nature of decisions, the theory of decision-making processes should be of holistic approach, underlying the role of different elements and processes in decision-making activities. As Krogerus and Tschäppeler (2011) state, there are different tools and methods that may support the decision-making process. One of them is the Eisenhower Matrix, also known as the Urgent-Important-Matrix, facilitating the decisions when the tasks should be done, according to their importance. Thus, issues to be done are divided into the ones that are urgent and have to be done or delegated and the ones that are not urgent and one has to decide when they will be done or delete them. Another technique used in management is SWOT analysis which is applied to evaluate the project's strengths, weaknesses, opportunities and threats. In addition, costs and benefits can be identified by using the BCG Box. The Boston Consulting Group developed a method to estimate investments by using the concepts of cash cows, stars, question marks and dogs. Moreover, Maslow's theory of human needs offer information on the priority on making decisions. Decisions differ when one takes into account the type of data; the ones made on the basis of a relatively small amount of data are different from the ones that have to be made in the face of big data.

Consequently, decisions involving big data may involve the participation of machines that facilitate data gathering and comprehension. Moreover, decisions involving big data are often supported by statistical or econometric tools. In making decision, the characteristics of big data are crucial. As Firican (2017) mentions, there are ten Vs of Big Data. They include: volume, velocity, variety, variability, veracity, validity, vulnerability, volatility, visualization, and value. Big data can, among others, help understand the decisions made by customers, target their needs and expectations as well as optimize business processes. In addition, Podnar (2019) draws our attention to the issue of new data privacy regulations. Thus, the company or the individual gathering information should identify sensitive data and treat it in the proper way, e.g., by using special encryption. Moreover, all processes connected with data, that is gathering, storing, modifying and erasing should be done according to the legal regulations, e.g., the EU's General Data Protection Regulation (GDPR).

Decision Theory and Game Theory

Although the first traces of game theory can be noticed in the works of economists in the previous centuries, the origin of game theory is associated with the book by John Neumann and Oskar Morgenstern *Theory of Games and Economic Behavior* printed in 1944. In the 1950's John Forbes Nash Jr. published papers on non-cooperative games and the Nash equilibrium. Since the publication of seminal works by Neumann, Morgenstern and Nash, the findings of game theory started to be applied in disciplines related to mathematics as well as the ones not connected with it. The application of game theory includes different spheres of study, such as economics, linguistics, translation, biology, anthropology, to mention a few of them. Professor Nash was the Nobel Prize winner in economics in 1994 for the work on the game theory. Apart from the mentioned scientific recognition of this theory among researchers, the interest in the game theory is also connected with the general curiosity shared by both researchers and laymen about how people make choices and what drives their selection.

The growing role of decision-making in both professional and private life has led to the increasing popularity of game theory in various scientific disciplines as well as in the studies representing how individuals behave in everyday situations. In addition, games as such experience their renaissance in the twenty-first century, being present in different spheres of life, not exclusively only the ones related to entertainment. The development in the sphere of games is also connected with new types of technologically advanced appearing on the market. Moreover, the growing role of technology and the Internet resulted in novel forms of competition and cooperation. Apart from the vivid interest among researchers representing different domains in the possibilities of applying game theory to study the nuances of a given discipline, game theory is known to a wider public because of the biographical drama film entitled *A Beautiful Mind* directed by Ron Howard showing the life of Professor John Nash.

Game Theory – Definition and Basic Concepts

As far as the definition of game theory is concerned, Philip D. Straffin in his book states that game theory examines the logical analysis of conflict and cooperation. Thus, the concept of a game is used if there are at least two players (human beings and non-human beings, such as communities, companies or countries) involved in cooperative and conflicting activities. Although games concern mainly human beings or companies, they can also be studied during the observation of plants and animals. Moreover, every player has some strategies at his or her disposal that can be used to play a game. The combination of strategies selected by players determines the result of a game. The outcome is connected with the payoff for players that can be exemplified in numbers. Game theory investigates how players should play in a rational way leading to the highest possible payoffs. The results of a game are determined by the choices made by a player and other players. The decisions of other players can be studied through the perspective of cooperation and conflict. Conflict is connected with different needs of

players, often having contradictory aims. Cooperation reflects the situation when the coordination of interests takes place. Although game theory can be applied to many situations and players, Straffin draws one's attention to the potential limitations of game theory. First, games played in a real world are complicated, with the total number of players and the outcomes of their strategies difficult to estimate. The second challenge is connected with the assumption that a player behaves in a rational way. In reality, not all players perform rationally, some behaviors cannot be easily explained. The third problem is connected with the fact that game theory cannot predict how the game evolves if the aims of players are not contradictory in a distinguishable way or when more than two players take part in a game. For such games, partial solutions, cases and examples exist. Thus, some games fail easy categorization and additional research has to be carried out to understand the complex picture of a given phenomenon. Games can be classified taking into account the number of players, types of payoffs and potential outcomes. Starting with the last feature, zero-sum games encompass situations with completely antagonistic aims, when one wins and the other loses. On the other hand, in non-zero-sum games the winning of one player does not necessarily entail losing of the other one. The taxonomy of games related to players includes the division of games according to the number of players (from one-to-many users). Games can also be classified through the prism of signaling. John Neumann and Oskar Morgenstern highlighted that *inverted signaling*, aimed at misleading the other player, can be observed in most games. *Direct signaling*, on the other hand, takes place very rarely in games. The payoffs in games depend on, among others, type of game and discipline it is applied in. They can take the form of money or utility (e.g., economics) as well as fitness from the genetic perspective (biology).

Game Theory Strategies, Decision Theory and Big Data

Elvis Picardo describes basic game strategies. One of them is *Prisoner's Dilemma* that shows how

acting in one's own interests leads to worse outcomes when a cooperation is chosen. In the *Prisoner's Dilemma* two suspects of a crime are detained in separate rooms, without the possibility of communicating with each other. Each of them is informed that if he/she cooperates and testifies against the second detainee, he or she will go free. When he or she decides not to cooperate but the other prisoner opts for cooperation, he or she will have to spend 3 years in prison. When both prisoners decide to confess, they will be imprisoned for 2 years. If none of them cooperates they will spend 1 year in prison. Although cooperation is the best selection for both prisoners, the most often chosen option is confessing against the other participant. Picardo in his contribution also shows more advanced game theories that rely on the Prisoner's Dilemma. One of them is *Matching Pennies* in which two players place a penny simultaneously on the table, with payoffs depending how often heads or tails appear. If both coins are heads or tails, the first player wins and can take the second player's coin. When one penny turns heads and the other tails, the second player is the winner. A similar social choice to the one of the Prisoner's Dilemma is represented in *Deadlock*, with the dominant strategy being the selection of the greatest benefit for both sides. Another type of advanced game theory is *Cournot Competition*, used in, e.g., depicting such economic phenomena as duopoly. An example of sequential game is *Centipede Game*, with players making moves one after another.

Cross-References

- [Economics](#)
- [Knowledge Management](#)
- [Online Advertising](#)
- [Social Network Analysis](#)

Further Reading

- Bielenia-Grajewska, M. (2013). International neuromanagement. In D. Tsang, H. H. Kazeroony, & G. Ellis (Eds.), *The Routledge companion to international management education*. Abingdon: Routledge.

- Bielenia-Grajewska, M. (2014). CSR Online Communication: the metaphorical dimension of CSR discourse in the food industry. In R. Tench, W. Sun, & B. Jones (Eds.), *Communicating corporate social responsibility: perspectives and practice (Critical studies on corporate responsibility, governance and sustainability)* (Vol. 6). Bingley: Emerald Group Publishing Limited.
- Bielenia-Grajewska, M. (2015). Neuroscience and learning. In R. Gunstone (Ed.), *Encyclopedia of science education*. New York: Springer.
- Bielenia-Grajewska, M. (2016). Good health is above wealth. Eurozone as a patient in eurocrisis discourse. In N. Chitty, L. Ji, G. D. Rawnsley, & C. Hayden (Eds.), *Routledge handbook on soft power*. Routledge: Abingdon.
- Firican, G. (2017). The 10 Vs of big data. Available online: <https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx>. Accessed June 2019.
- Krogerus, M., & Tschäppeler, R. (2011). *The decision book: fifty models for strategic thinking*. London: Profile Books Ltd..
- Picardo, E. (2016). Advanced game theory strategies for decision-making. Investopedia. <http://www.investopedia.com/articles/investing/111113/advanced-game-theory-strategies-decisionmaking.asp>. Accessed 10 Sept 2016.
- Podnar, K. (2019). How to survive the coming data privacy Tsunami. Available at <https://tdwi.org/Articles/2019/06/17/DWT-ALL-How-to-Survive-Data-Privacy-Tsunami.aspx>. Accessed 20 June 2019.
- Straffin, P. G. (2004). *Teoria Gier (Game theory and strategy)*. Warszawa: Wydawnictwo Naukowe Scholar.
- Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Woodstock: Princeton University Press.

Deep Learning

Rayan Alshamrani and Xiaogang Ma
Department of Computer Science,
University of Idaho, Moscow, ID, USA

Introduction

Artificial intelligence (AI) is a growing and well-known discipline in computer science with essentially specialized fields. Deep learning is a major part of AI that is associated with machine learning notion. As a branch of machine learning, deep learning focuses on approaches to improve AI systems through learning and training from experience and observation. Deep learning relies

on a variety of studies such as applied mathematics, statistics, neuroscience, and human brain knowledge. The prime principles of applied mathematics, such as linear algebra and probability theories, are the major disciplines that inspire the fundamentals of the modern deep learning. The main idea behind deep learning is to represent real-world entities as related concepts of a nested hierarchy, where each concept is defined by its relation to a simpler concept. Therefore, deep learning aims to empower computers to learn from experience and to understand different domains regarding a specific hierarchy of concepts. This allows computers to build and learn complex concepts and relationships from gathering the related simpler concepts.

Deep learning began to gain its popularity in the middle 2000s. At that time, the initial intention was to make more generalized deep learning models with small datasets. Today, deep learning models have achieved great accomplishments by leveraging large datasets. Moreover, a continuous accomplishment in deep learning is the increase of model size and performance due to the advances in general-purpose CPUs, software infrastructure, and network connectivity. Another successful achievement in today's deep learning is its enhanced ability to make predictions and recognitions with high level of accuracy, unambiguity, and reliability. Deep learning's ability to perform tasks with high-level complexity is escalating. Thus, many modern applications have successfully applied deep learning from different aspects. Furthermore, deep learning provides useful tools to process massive amounts of datasets and makes practical contributions to many other scientific domains.

The History of Deep Learning

Through the history, computer science scholars have defined deep learning by different terms that reflect their different points of view. Although many people assumed that deep learning is a new discipline, it has existed since the 1940s, but it was not quite popular back then. The evolution of deep learning started between the 1940s

and the 1960s when the researchers knew it as *cybernetics*. Between the 1980s and 1990s, deep learning scientists acknowledged it as *connectionism*. The current name of this discipline, *deep learning*, took its shape around the first decade of the 2000s up until today. Hence, the previous terminologies illustrate the revolution of deep learning through three different waves (Goodfellow et al. 2017).

Deep Learning as a Machine Learning Paradigm

The main idea behind machine learning is to consider generalization when representing the input data and to train the machine learning models with these sets of generalized input data. By doing so, a trained model is able to deal with new sets of input data in future uses. Hence, the efficient generalization of the data representation has a huge impact on the performance of the machine learners. However, what will happen if these models generate unwanted, undesired, or incorrect results? The answer to this question is to feed these models with more input data. This process forms a key limitation in machine learning. Besides, machine learning algorithms are limited in their ability to perform on and extract raw forms from natural data. Because of this, machine learning systems require considerable domain expertise and a high level of engineering in order to design models that extract raw data and transform it into useful data representation.

As mentioned earlier, deep learning relies on the hierarchical architecture where lower level features define high level features. Because of their nature, deep learning algorithms aid agents to overcome the machine learning algorithms' limitations. Deep learning algorithms support machine learners by extracting data representation with a high level of complexity. This data extraction mechanism feeds machine learners with raw data and enables these learners to automatically discover the suitable data representations. Deep learning is beneficial for machine learning because it enables machine learners to handle a large amount of input datasets, especially unsupervised datasets. Consequently, deep

learning algorithms yield better and promising results in different machine learning applications such as computer vision, natural language processing, and speech recognition. Deep learning is an important achievement in AI and machine learning. It enhances the agents' abilities to handle complex data representations and to perform AI tasks independently from human knowledge. In summary, deep learning introduces several benefits which are: (1) enabling simple models to work with knowledge acquired from huge data with complex representations, (2) automating the extraction of data representations which makes agents work with different data types, and (3) obtaining semantic and relational knowledge from the raw data at the higher level of representations.

Deep Learning Applications and Challenges

Since the 1990s, many commercial applications have been using deep learning more as a state-of-the-art concept than applied technology, because the comprehensive application of deep learning algorithms need expertise from several disciplines, and only a few people were able to do that. However, the number of skills required to cope with today's deep learning algorithms is in decrease due to the availability of a huge amount of training datasets. Now, deep learning algorithms and models can solve more complicated tasks and reach high-level human performance.

Deep learning can perform accurate and valid tasks, such as prediction and recognition, with a high level of complexity. For instance, nowadays deep learning models can recognize objects in photographs without the need to crop or resize the photograph. Likewise, these models can recognize a diversity of objects and classify them into corresponding categories. Besides object recognition, deep learning also has some sort of influence on speech recognition. As deep learning models drop the error rate, they could recognize voices more accurately. Traffic sign categorization, pedestrian detection process, drug discovery, and image segmentation are examples of deep

learning's recent successful case studies. Accordingly, many companies such as Apple, Amazon, Microsoft, Google, IBM, Netflix, Adobe, and Facebook have increased their attention towards deep learning as they are positively profitable in business applications.

In contrast, with these successful achievements come the drawbacks and limitations. There are major challenges associated with deep learning that remain unsettled and unresolved, especially when it comes to big data analytics. First, there are specific characteristics that cause the drawbacks and limitations of adopting deep learning algorithms in big data analytics, such as models' scalability, learning with streaming data, distributed computing, and handling high dimensional data (Najafabadi et al. 2015). Second, the nature of deep learning algorithms, which briefly map objects through a chain of related concepts, inhibits it from performing commonsense reasoning exercises regardless of the amount of data being used. Third, deep learning has some limitations when performing classification on unclear images due to the imperfection of the model training phase. This imperfection makes the potential deep learning model more vulnerable to the unrecognizable data. Nevertheless, several research contributions are validating and embracing techniques to improve the deep learning algorithms against the major limitations and challenges.

Concepts Related to Deep Learning

There are several important concepts related to deep learning, such as reinforcement learning, artificial neural networks, multilayer perception, deep neural networks, deep belief networks, and backpropagation.

A key success in deep learning is its extension to the reinforcement learning field. Reinforcement learning helps an agent to learn and observe through trial and error without any human intervention. The existence of deep learning has empowered reinforcement learning in robotics. Moreover, reinforcement learning systems that apply deep learning are performing tasks at human level. For example, these systems can

learn to play Atari video games just like professional gamers.

In order to fully understand deep learning basics, the concept of artificial neural networks (ANN) must be illustrated. The main idea of the ANN is to demonstrate the learning process of the human brain. The structure of the ANN consists of interconnected nodes called neurons and a set of edges that connect these neurons all together. The main functionality of ANN is to receive a set of inputs, perform several procedures (complex calculations) on input sets, and use the resulted output to solve specific real-world problems. ANN is highly structured with multiple layers, namely the input layer, the output layer, and the hidden layer in between.

Multilayer perception (MLP), also called feedforward neural networks, or deep feedforward networks, is a workhorse of deep learning models. MLP is a mathematical function (called function f^*) that maps input to output. This f^* function is formulated through composing several simple functions. Each of these simple functions provides a new way to represent the input data. Deep learning models that adopt MLP are known as feedforward because the information flows from the input to the output through the model's function without any feedback.

Broadly speaking, a deep neural network (DNN) learns from multiple and hierarchical layers of sensory data representation, which enables it to perform tasks at a level close to human ability. This makes DNNs more powerful than shallow neural networks. DNN layers are divided into early layers, which are dedicated for identifying simple concepts of input data, and later layers, which are dedicated for complex and abstract concepts. The DNN differs from the shallow neural network in the number of hidden layers. A DNN has more than two hidden layers. With that said, a network is deep if there are many hidden layers.

Deep belief networks (DBN) is a type of DNN. Specifically, DBN is a generative graphical model with multiple layers of stochastic latent variables consisting of both directed and undirected edges. The multiple layers of DBN are hidden units. DBN layers are connected with each other, but units within each layer are not. Namely, DBN is a

stack of Restricted Boltzmann Machine, and it uses Greedy Layer-wise algorithm for model training.

Backpropagation is a quintessential supervised learning algorithm for various neural networks. Backpropagation algorithm is a mathematical tool, which computes weights' gradient descent, for improving predictions accuracy. The aim of the backpropagation algorithm is to train neural networks by comparing the initial output with the desired output and then adjust the system until the comparison difference is minimized.

Future of Deep Learning

Deep learning will have a very bright future with many successes as the world is moving toward the big data era with the rapid growth in the amount and type of data. The expectation is on the rise because this valuable discipline requires very little engineering by manual work, and it leans heavily on the automation of data extraction, computation, and representation. Ongoing research have outlined the need to combine different concepts of deep learning together to enhance applications such as recognition, gaming, detection, health monitoring, and natural language processing. Thus, the evaluation of deep learning should focus on working more with unsupervised learning, so agents can mimic the human brain behavior further and start thinking on behalf of human beings. Researchers will continue introducing new deep learning algorithms and forms of learning in order to develop general purpose models with high levels of abstraction and reasoning.

The big data bang poses several challenges to deep learning, such as those listed above. Indeed, most big data objects consist of more than one modality, which require advanced deep learning models that can extract, analyze, and represent different modalities of input datasets. It is true that big data offers satisfiable amount of datasets to train the deep learning models and improve their performance. Yet, the process of training deep learning models using huge datasets depends significantly on high-performance computing infrastructure, which is sometimes challenging

due to the fact that the growth rate of big data is faster compared to the gain in computational performance (Zhang et al. 2018). However, many recent studies have outlined deep learning models that are suitable for big data purposes. It is noticeable that deep learning algorithms have made great progress in big data era, and the challenges that face deep learning in big data era are under current and prospective research consideration.

Emerging semantic technologies with deep learning is a cutting-edge research topic. This emergence has created the notion of semantic deep learning. For instance, the key successes of both semantic data mining and deep learning have inspired researchers to potentially assist deep learning by using formal knowledge representations (Wang 2015). Similarly, deep learning approaches would be beneficial for evaluating semantic similarity of two sentences with 16–70% improvement compared to baseline models (Sanborn and Skryzalin 2015). In addition to semantic technologies, Decision Support Systems (DSS) is another aspect that will gain more advantages from the adaptation of deep learning. The current studies that relate DSS to deep learning focus more on applying deep learning concepts to the clinical DSS in healthcare. It is possible to see more DSSs in different domains that use deep learning methods in the near future. Deep learning is an important part of machine learning. As most researchers are looking for ways to simulate the biological brain, deep learning will be powerfully presented in machine learning studies and applications.

Cross-References

- [Artificial Intelligence](#)
- [Deep Learning](#)

Further Reading

- Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning* (Vol. 1). MIT press.
- Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep

- learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 1.
- Sanborn, A., & Skryzalin, J. (2015). Deep learning for semantic similarity. *CS224d: Deep Learning for Natural Language Processing Stanford, CA, USA: Stanford University*.
- Wang, H. (2015). Semantic Deep Learning. *University of Oregon*, 1–42.
- Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, 42, 146–157.

Deep Web

- [Surface Web vs Deep Web vs Dark Web](#)

Defect Detection

- [Anomaly Detection](#)

De-identification

- [Anonymization Techniques](#)

De-identification/Re-identification

Patrick Juola
Department of Mathematics and Computer Science, McNulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Introduction

Big data often carries the risk of exposing important personal information about individuals in the database. To reduce the risk of such privacy violations, databases often mask or anonymize individual identifiers in databases, a process known as “*de-identification*.” Unfortunately, it is often possible to infer the identities of individuals from

information remaining in the database and thus to “*re-identify*” them. The development of robust and reliable methods of de-identification is an important public policy question in the handling of big data.

Privacy and Personally Identifiable Information

Databases often contain information about individuals that can be embarrassing or even harmful if widely known. For example, university records might reveal that a student had failed a class several times before passing it; hospital records might show that a person had been treated for an embarrassing disease, and juvenile criminal records might show arrests for long-forgotten misdeeds, any of which can unjustly harm an individual in the present. However, this information is still useful for researchers in education, medicine, and criminology provided that no individual person is harmed. De-identifying this data protects privacy-sensitive information while allowing other useful information to remain in the database, to be studied, and to be published.

The United States (US) Department of Health and Human Services (HHS 2012), for example, references 19 types of information (Personally Identifiable Information, or PII) that should be removed prior to publication. This information includes elements such as names, telephone, and fax numbers, and biometric identifiers such as fingerprints. This list is explicitly not exhaustive, as there may be another “unique identifying number, characteristic, or code” that must also be removed, and determining the risk of someone being able to identify individuals may be a matter of expert judgment (HHS 2012).

De-identifying Data

By removing PII from the database, the presumption is that the remaining de-identified information no longer contains sensitive information and therefore can be safely distributed. In addition to simply removing fields, it is also possible to adjust the data delivery method so that the data delivered

to analysts does not allow them to identify individuals. One such method, “statistical disclosure limitation,” masks the data by generating synthetic (meaning, “fake”) data with similar properties to the real data (Rubin 1993). Another method (“differential privacy”) is to add or subtract small random values to the actual data, enough to break the link between any individual data point and the person it represents (Garfinkel 2015).

However, de-identification brings its own problems. First, as HHS recognizes, “de-identification leads to information loss which may limit the usefulness of the resulting health information in certain circumstances.” (HHS 2012). Other researchers agree. For example, Fredrikson (2014) showed that using differential privacy worsened clinical outcomes in a study of genomics and warfarin dosage. More seriously, it may be possible that the data can be re-identified, defeating the purpose of de-identification.

Re-identification

Even when PII is removed, it may be possible to infer it from information that remains. For example, if it is known that all patients tested in a given month for a specific condition were positive, then if a person knows that a particular patient was tested during that month, the person knows that patient tested positive.

A common way to do this is by using one set of data and linking it to another set of data. In one study (Sweeney 2000), a researcher spent \$20 on a set of voter registration records and obtained the ZIP code, birth date, and gender of the Governor of Massachusetts. She was then able to use this to identify the Governor’s medical records via a public Federal database. She estimated that more than 85% of the US population can be identified uniquely from these three datapoints. Even using counties, instead of the more informative ZIP codes, she was able to identify 18.1% of the US population uniquely. It is clear, then, that re-identification is an issue even when all the obvious unique links to individuals have been purged (Sweeney 2000; Garfinkel 2015).

For this reason, de-identification and re-identification remain active research areas and should be treated with concern and caution by any analyst dealing with individual human data elements or other sensitive data.

Cross-References

► Profiling

Further Reading

- Fredrikson, M., et al. (2014). *Privacy in pharmacogenetics: An end-to-end case study of personalized Warfarin Dosing*. 23rd Usenix Security Symposium, August 20–22, 2014, San Diego, CA.
- Garfinkel, S. L. (2015). *De-identification of personal information*. NISTIR 8053. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.IR.8053>.
- Health and Human Services (HHS), US. (2012). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule*. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveridentities/De-identification/hhs_deid_guidance.pdf.
- Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2), 461–468.
- Sweeney, L. (2000). *Simple demographics often identify people uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh. <http://dataprivacylab.org/projects/identifiability/paper1.pdf>.

Demographic Data

Jennifer Ferreira

Centre for Business in Society, Coventry University, Coventry, UK

The generation of demographic data is a key element of many big data sets. It is the knowledge about people which can be gleaned from big data which has the potential to make these data sets even more useful not only for researchers but also policy makers and commercial enterprises. Demography, formed from two Greek words,

broadly means “description of the people” and in general refers to the study of populations, processes, and characteristics including population growth, fertility, mortality, migration, and population aging, while the characteristics examined are as varied as age, sex, birthplace, family structure, health, education, and occupation. Demographic data refers to the information that is gained about these characteristics which can be used to examine changes and behaviors of population and in turn be used to generate population predictions and models. Demographic data can be used to explore population dynamics, analytical approaches to population change, the demographic transition, demographic models, spatial patterns, as well as planning, policy making, and commercial applications especially projecting and estimating population composition and behavior.

Applied demography seeks to emphasize the potential for the practical application of demographic data to examine present and future demographic characteristics, across both time and space. Where demographic data is available over time, this allows for historical changes to populations to be examined in order to make predictions and develop models about how populations may behave in the future. Traditionally the principal sources for the study of population are censuses and population surveys which are often infrequent and not always comprehensive.

Understanding demographic data and population patterns has useful applications in many areas, indulging planning, policy making, and commercial enterprises. In planning, estimates and projections are important in terms of ensuring accurate allocation of resources according to population size, determining the level of investment needed for particular places. Planning requires reliable demographic data in order to make decisions about future requirements, and so will impact on many major planning decisions, and therefore many large financial decisions are made on the basis of demographic data in combination with other information. Population statistics reveal much about the nature of society, the changes that take place within it, and the issues that are relevant for government and policy. Therefore, demographic projections and models

can also stimulate actions in policy making, in terms of how to meet the needs of present and future populations. A key example of this relates to aging populations in some developed countries where projections of how this is likely to continue have informed policies around funding pensions and healthcare provision for the elderly. For businesses, demographic data is a vital source of information about their consumer base (or potential consumer base); understanding how consumers behave can inform their activities or how they target particular cohorts or segments of the population. The wide relevance of demographic data for policy, planning, research, and commerce means that this particular aspect of big data has experienced much attention.

Many of the traditional data sets which generate demographic data, such as the census, are not conducted frequently (every 10 years in the UK) and are slow to release the results, so any analysis conducted on a particular population will often be significantly out of date given the constant changing dynamics of human populations. Furthermore, other population surveys draw on a relatively small sample of the population and so may not be truly representative of the range of situations experienced in a population.

Demographic data refers to data which relates to a particular population which is used to identify particular characteristics or features. There are a wide range of demographic variables which could be included in this category, although these commonly used include age, gender, ethnicity, health, income, and employment status. Demographic data has been a key area which has been the focus for big data research, primarily because much of the data generated relates to individuals and therefore has the potential to provide insights into the characteristics and behaviors of population beyond what is possible from traditional demographic data sources. Demographic data is also vital for many commercial enterprises as they seek to explore the demographic profile of their customer base or the customer base whom they wish to target for their products.

This vast new trove of big data generated by new technological advancements (mobile phone, computers, satellites, and other electronic

devices) has the potential to transform spatial-temporal analyses of demographic behavior, particularly related to economic activity. As a result of the technological advancements which have led to the generation of many big data sets, the quantity of demographic data available for population research is increasing exponentially. Where there are consistent large-scale data sets that now extend over many years sometimes crossing national boundaries with fine geographic detail, this collectively creates a unique laboratory for studying demographic processes and for examining social and economic scenarios. These models are then used in order to explore population changes including fertility, mortality, and depopulation.

The growth in the use of big data in demographic research is reflected in its growing presence in discussions at academic conferences. In the Population Association for America in 2014, a session entitled “Big Data for Demographic Research” demonstrates some of the ways big data has been used.

Mobile phone data in Estonia was used to examine ethnic segregation. This data set included information about the ethnicity of individuals (Russian/Estonian), the history of locations visited by the individuals, and their phone-based interactions. This study found evidence to suggest that ethnic composition of an individual’s geographic neighborhood influenced the structure of an individual’s geographic network. It also found that patterns of segregation were evident where migrants were more likely to interact with other individuals of their ethnicity. A further study also used mobile phone data to explore human mobility. This project highlighted the potential that large-scale data sets like this have for studying human behavior on a scale not previously possible. It argued that some measures of mobility using mobile data are contaminated by infrastructure and demographic and social characteristics of a population. The authors also highlight problems with using mobile phone data to explore mobility and outline potential new methods to measure mobility

in ways which respond to these concerns. The measures developed were designed to address the spatial and social nature of human mobility, to remain independent of social, economic, political, or demographic characteristics of context, and to be comparable across geographic regions and time.

The generation of big data via social media has also led to the development of new research methods. Pablo Mareos and Jorge Durand explore the potential value of netnographic methods in social media for migration studies. To do this the researchers explore data obtained from Internet discussion forums on migration and citizenship. This research uses a combination of classification methods to analyze discussion themes of migration and citizenship. The study revealed results which identified key migrating practices which were absent from migration and citizenship literature, suggesting that analyses of big data may provide new avenues for research, with the potential for this to revolutionize traditional population research methods.

Capturing demographic patterns from big data is a key activity for researchers, political teams, and marketing teams. Being able to examine the behavior of populations, but in particular specific segments of population is a key activity. While many of the techniques and technologies which harness big data may have been developed by commercial enterprises or for commercial gain, there are an increasing number of examples of where this data is being used for public benefit, as seen in Chicago.

In Chicago, health officials employed a data analytics firms to conduct data mining to explore ethnic minority women who were not getting breast screenings even though they were offered them for free at a hospital in a particular area. The analytics firm helped Chicago health department to refine its city outreach for breast cancer screening program by using big data to identify the uninsured women aged 40 and older living in the south side of the city. This project indicates the potential impact that big data could have on public services.

There are of course challenges associated with using demographic data collected in data sets. The use of Twitter data, for example, to examine population patterns or trends though is problematic in that population of Twitter users is not necessarily representatives of the wider population or the population being studied. The massive popularity of social media, and the ability to extract the data about communication behaviors, has made them a valuable data source. However, a study which compared the Twitter population to the US population along three axes (geography, gender, and race/ethnicity) found that the Twitter population is a highly non-uniform sample of the population. Ideally when comparing the Twitter population to society as a whole, we would compare properties including socioeconomic status, education level, and type of employment. However, it is only possible to obtain characteristics which are self-reported and made visible by the user in the Twitter profile (usually the name, location, and text included in the tweet). Research has indicated that Twitter users are more likely to live within densely populated area and that sparsely population regions are underrepresented. Furthermore, research suggested that there is a male bias in Twitter users, again making the results gained from this big data source unrepresentative of the wider population.

Despite the challenges and limitations associated with big data, the study of populations and their characteristics and behaviors is a growing area for big data researchers. The applications for demographic data analysis are being adopted and explored by data scientists in both the public and private sector in an effort to explore the past, present, and future patterns of populations across the world.

Cross-References

- [Education](#)
- [Epidemiology](#)
- [Geography](#)

Further Reading

- Blumenstock, J., & Toomet, O. (2014). *Segregation and 'Silent Separation': Using large-scale network data to model the determinants of ethnic segregation*. Paper presented at the population association of America 2014 annual meeting, Boston, May 1–3.
- Giroi, F., & King, G. (2008). *Demographic forecasting*. Princeton/Oxford: Princeton University Press.
- Mateos, P., & Durand, J. (2014). *Netnography and demography: Mining internet discussion forums on migration and citizenship*. Paper presented at the population association of America 2014 annual meeting, Boston, 1–3 May.
- Mislove, A., Lehmann, S., Ahn, Y. -Y., Onnela, J. -P., & Rosenquist, N. Understanding the demographics of Twitter users. In *Proceedings of the fifth international AAAI conference on weblogs on and social media*. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234>.
- Ruggles, S. (2014). Big microdata for population research. *Demography*, 51(1), 287–297.
- Rowland, D. (2003). *Demographic methods and concepts*. Oxford: Oxford University Press.
- Sobek, M., Cleveland, L., Flood, S., Hall, P., King, M., Ruggles, S., & Shroeder, M. (2011). Big data: Large historical infrastructure from the Minnesota population center. *Historical Methods*, 44(2), 61–68.
- Williams, N., Thomas, T., Dunbar, M., Eagle, N., & Dobra, A. (2014). *Measurement of human mobility using cell phone data: Developing big data for demographic science*. Paper presented at the population association of America 2014 annual meeting, Boston, 1–3 May.

Digital Advertising Alliance

Siona Listokin

Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

The Digital Advertising Alliance (DAA) is a non-profit organization in the United States (US) made up of marketing and advertising industry associations that seeks to provide self-regulatory consumer privacy principles for internet-based advertising. The DAA is one of the most prominent self-regulation associations in consumer data privacy and security but has been criticized for promoting weak data privacy programs and enforcement.

The DAA was established in 2009 by several US advertising associations, following the release of a Federal Trade Commission (FTC) report on “Self-Regulatory Principles for Online Behavioral Advertising.” It is led by the Association of National Advertisers, The American Advertising Federation, 4A’s, Network Advertising Initiative, Better Business Bureau National Programs, and Interactive Advertising Bureau. The DAA represents thousands of advertising and marketing companies and includes hundreds of participating companies and organizations, across a range of industries. Originally, participating companies consisted of advertisers and third party analytics companies, but starting in 2011, DAA expanded its efforts to include social networks and non-advertising firms.

The Alliance’s major self-regulatory guidelines stem from its “Principles for Internet Based Advertising” released in mid-2009 and form the basis for the DAA AdChoices icon and the consumer opt-out program for customized ads. The alliance has issued applications of its principles to digital advertising areas including political ads, cross-device data use, mobile, multisite data, and online behavioral advertising. The self-regulatory principles, which participating companies can highlight with the DAA’s blue icon, are administered by the Advertising Self-Regulatory Council (ASRC) of the Council of Better Business Bureaus and the Association of National Advertisers. The principles are explicitly meant to correspond with the FTC’s report and focus on consumer education, transparency, control, data security, consent, and sensitive data like health, financial, and child-directed data. The DAA icon, launched in October 2010, is meant to serve as a signaling device to consumers that informs users of tracking activities.

The DAA’s consumer opt-out page, known as “Your AdChoices,” is an element of the icon program that allows users to click in and choose to opt-out of specific interest-based advertising. The page formed in November 2010 as the Alliance participated in, and subsequently withdrew from, the World Wide Web Consortium’s working group on “Do Not Track” standardization. Consumers can visit the opt-out page and select to opt-out of participating third parties’

browser-enabled personalized advertising. DAA created a separate App Choices program for consumer control of mobile app data collection. While the opt-out option applies to behavioral advertising, data collection and third party tracking are not blocked.

Enforcement and Criticism

Enforcement is handled by the Association of National Advertisers (ANA) and the Better Business Bureau National Programs (BBBNP); DAA refers to their independent enforcement component though it is worth noting that these organizations are participating and founding associations. In cases of potential noncompliance, the BBBNP’s Digital Advertising Accountability Program (DAAP) process begins. DAAP sends an inquiry letter to the company and may begin a formal review with subsequent decision. Since 2011 (through the first half of 2020), the BBBNP has ruled on a total of 80 cases through DAAP. In the first half of 2020, the ANA received about 4000 consumer inquiries about online advertising, repeating 2019s jump in consumer complaints in this area from a previous average of about 500 a year. These inquiries range from concerns about ads blocking web content, indecent advertisements, and incorrectly targeted ads. In rare occasions, the alliance refers a case to the FTC, and a number of FTC Commissioners have supported DAA oversight in speeches and reports.

The DAA has been criticized by advocacy groups and policymakers for failing to provide meaningful privacy protection and transparency to consumers. Although the DAA had agreed to the Do Not Track effort in principle after the FTC recommended it, it disagreed with the extent of tracking restrictions proposed by the working group and declared that it would not penalize advertisers that ignore the standards. The DAA’s own AdChoices opt-out relies on cookies that must be manually updated to prevent new third party tracking and that can negatively impact user experience. In 2013, Senator John D. Rockefeller IV criticized the DAA’s opt-out program for having too many exceptions that allow for consumer tracking for market research. A 2018 study noted

major usability flaws in the mobile app opt-out program.

Further Reading

- Federal Trade Commission. (2009). *FTC staff report: Self-regulatory principles for online behavioral advertising, 2009*. Washington, DC: Federal Trade Commission.
- Garlach, S., & Suthers, D. (2018). *I'm supposed to see that? 'AdChoices usability in the mobile environment*. Proceedings of the 51st Hawaii international conference on System Sciences.
- Mayer, J. R., & Mitchell, J. C. (2012). *Third-party web tracking: Policy and technology*. Security and Privacy (SP), 2012 IEEE Symposium on IEEE.
- Villafranco, J., & Riley, K. (2013). So you want to self-regulate? The national advertising division as standard bearer. *Antitrust*, 27(2), 79–84.

Digital Agriculture

► Agriculture

Digital Divide

Lázaro M. Bacallao-Pino
University of Zaragoza, Zaragoza, Spain
National Autonomous University of Mexico,
Mexico City, Mexico

Synonyms

Digital inequality

The notion of the “digital divide” came into broad use during the mid-1990s, beginning with reports about access to and usage of the Internet published by the United States (US) Department of Commerce, National Telecommunications and Information Administration in 1995, 1998, and 1999. The term was defined as the divide between those with access to information and communication technologies (ICTs) and those without it and was considered one of the leading economic and civil rights issues in contemporary societies. Since

then, the digital divide has rapidly and widely become a topic of research for both policymakers and scholars, calling attention to the problem of unequal access to ICTs. That unequal access also raises questions of big data in relation to concerns about tracking and tracing usage and privacy and of the implications of accelerated data flows to already data-rich as opposed to data-poor contexts for widening digital divides.

From the beginning of debates on the digital divide, there have been different positions regarding the possibility of overcoming gaps in access between different countries, groups, and individuals. On the one hand, related digital inequalities have been framed as a temporary problem that will gradually fade over time due to two factors: steadily decreasing costs of use of the Internet and its continuously increasing ease of use. Based on these assumptions, some views have it that, instead of a source of divide, the Internet provided a technological opportunity for information freedom and, above all, a tool for illiterate people to learn and read, abridging what was considered the “real” divide – the gap between those who can read well and those who cannot – giving the latter opportunities to take advantage of easily accessible information resources. On the other hand, the digital divide has been considered a long-term pattern, generating a persistent division between “info-haves” and “info-have-nots.” In that sense, perspectives on the digital divide have distinguished among cyber-pessimists underlining deep structures and trends in social stratification that result in the emergence of unskilled groups without technological access; cyber-skeptics proposing a one-way interrelationship between society and technology in which the latter adapts to the former, not vice versa; and cyber-optimists proposing a positive scenario considering that, at least in developed countries, the digital divide will be bridged as a result of the combined action of technological innovations, markets, and state (Norris 2001).

As mentioned, from its original description, the digital divide was commonly defined as the perceived gap between those who have access to ICTs and those who do not, summarized in terms of a division between information “haves” and “have-nots.” From this perspective, it initially

was measured in terms of existing numbers of subscriptions and digital devices, but, as these numbers constantly increased and there was a transition from narrow-band Internet towards broadband DSL and cable modems in the early 2000s, more recently from an access perspective the digital divide has been measured in terms of the existing bandwidth per individual. It also is in this regard that increasingly massive data flows and collection from various devices have implications for tracking, monitoring, and measuring different dimensions or aspects of the digital divide. Since new kinds of connectivity are never introduced simultaneously and uniformly to society as a whole, level and quality of connectivity are associated with another degree of digital divide in terms of access.

Although access is important, at the same time, it has been noted that a binary notion of a “yes” or “no” proposition regarding physical access to computers or the Internet does not offer an adequate understanding of the complexity and multidimensionality of the digital divide. In this sense, technological gaps are related to other socioeconomic, cultural, ethnic, and racial differences, such that there is a necessity to rethink the digital divide and related rhetoric. This divide is relevant particularly in the context of what has been defined as the Information Age, such that not having access to those technologies and information is considered an economic and social handicap. Consequently, different approaches have aimed to broaden the notion of the digital divide to provide a more complex understanding of access, usage, meaning, participation, and production of digital media technology.

Besides transition towards a more complex and multidimensional point of view on the particularities of the inequalities created by the digital divide when comparing it to other scarce material and immaterial resources, one can also find efforts for understanding different types of access associated with the digital divide: from motivational and physical ones to others related to skills and usage. As part of the tendency towards this new

perspective on the digital divide, there has been a shift from an approach centered on physical access towards a focus on skills and usage, i.e., a second-level digital divide (Hargittai 2002).

The multidimensional nature of digital divide arguably refers to at least three levels: global, social, and democratic (Norris 2001). While the global divide is focused on the different levels of access between industrialized and developing countries, the social one refers to gaps among individuals who are considered as information-rich and poor in each country. At the democratic level, the digital divide is associated with the quality of the use of ICTs by individuals, distinguishing between those ones who use its resources for their engagement, mobilization, and participation in public life, and others who do not.

From access-focused trends, public policies to help bridge the digital divide have mainly focused on the development of infrastructures for providing Internet access to some groups. However, some researchers have judged the results of those actions as insufficient. On the one hand, and regarding physical access, many groups with low digital opportunities have been making substantial gains in connectivity and computer ownership. However, on the other hand, significant divides in Internet penetration persist between individuals, in close relationship to different levels of income and education, as well as other dimensions such as race and ethnicity, age, gender, disabilities, type of family, and/or geographic location (urban-rural). At the same time, while there may be trends in the digital divide closing in terms of physical access – mainly in the most developed countries – the digital divide persists or even widens in the case of digital skills and the use of applications associated with ICTs.

Differences in trends between physical access and digital skills show the complex interrelationships among the different levels at which the digital divide exists. Technical, social, and types of uses and abilities for effective and efficient use of ICTs are articulated in a multidimensional phenomenon in which the ways that people use

the Internet have rising importance for understanding the digital divide. Leaders in the corporate sector, governments and policymakers, nongovernmental organizations, and other civil society actors and social movements have been concerned about the digital divide, given the increasing centrality of the Internet to socialization, work, education, culture, and entertainment, as a source of training and educational advancement, information, job opportunities, community networks, etc.

In summary, from this point of view, especially in relation to civil society and commitments to social change, moving beyond a usage and skills-centered approach to the digital divide towards a perspective on the appropriation of digital technologies by socially and digitally marginalized groups involves articulation of both the uses of ITCs and meanings associate with them. Closing the digital divide is considered, from this perspective, part of a more general process of social inclusion, particularly in contemporary societies where the access to and creation of knowledge through ICTs are seen as a core aspect of social inclusion, given the rising importance of dimensions such as identity, culture, language, participation, and sense of community. More than overcoming some vision of the digital divide marked by physical access to computers and connectivity, considering it from an approach focused on technology for social inclusion and change, the digital divide is reoriented towards a more complex understanding of the effective articulation of ICTs into communities, institutions, and societies. It is in this regard that big data has been particularly engaged for measuring and analyzing the digital divide relative to processes of social development taking into account all the dimensions – economic, political, cultural, educational, institutional, and symbolic – of meaningful access to ICTs.

Cross-References

- [Cyberinfrastructure \(U.S.\)](#)
- [Digital Ecosystem](#)

- [Digital Literacy](#)
- [Information Society](#)

Further Reading

- Compaine, B. M. (2001). *The digital divide: Facing a crisis or creating a myth?* Cambridge, MA: MIT Press.
- Hargittai, E. (2002). Second-level digital divide: Differences in people's online skills. *First Monday*, 7(4). <https://doi.org/10.5210/fm.v7i4.942>.
- Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the internet worldwide*. Cambridge: Cambridge University Press.
- Van Dijk, J. A. G. M. (2006). Digital divide research, achievements and shortcomings. *Poetics*, 34(4–5), 221–235.
- Warschauer, M. (2004). *Technology and social inclusion: Rethinking the digital divide*. Cambridge, MA: MIT Press.

Digital Ecosystem

Wendy Chen

George Mason University, Arlington, VA, USA

The Definition of Digital Ecosystem

“Digital ecosystem” is a concept building upon “ecosystem” coined by British botanists Arthur Roy Clapham and Arthur George Tansley during the 1930 who argued that in nature, living organisms and the environment surrounding them interact with each other, which constitutes as an “ecosystem” (Tansley 1935). Since then, the ecosystem concept has been applied to various domains and studies including education and entrepreneurship etc. (Sussan and Acs 2017). Over the recent decades, with the rapid development of technology and internet, the “digital ecosystem” idea was born. It can be thought of as an extension of a biological ecosystem in the digital context, which relies on technical knowledge, and as “robust, self-organizing, and scalable architectures that can automatically solve complex, dynamic problems” (Briscoe and Wilde 2009).

The Applications of Digital Ecosystem

The digital ecosystem has been applied to an array of perspectives and areas in which it is studied including business, education, and computer science.

Business and Entrepreneurship

In business, digital ecosystem describes the relationship between a business and the end consumers in the digital world (Weill and Woerner 2015). New technology creates disruption for traditional business models. This process is referred to as *digital disruption*, such as books being read on eReaders rather than paperbacks or Uber's disruption of the taxi industry. A central theme in business for understanding digital ecosystems is for businesses to fully understand their end consumers, leveraging strong customer relationships, and increase cross-selling opportunities (Weill and Woerner 2015).

In entrepreneurship, the term Digital Entrepreneurship Ecosystem (DEE) refers to "an ecosystem where digital entrepreneurship emerges and develops" (Li et al. 2017). DEE reflects a group of entities that integrates resources to help facilitate and transform digital entrepreneurship (Li et al. 2017). In entrepreneurship literature, one of the fundamental differences between the digital entrepreneurial ecosystem and a traditional business ecosystem is the ventures in the digital entrepreneurial ecosystem focus on the interaction between digital technologies and users via digital infrastructure (Sussan and Acs 2017).

Education

In the education domain, teachers seek to expand the use of technology within the classroom to create a classroom digital ecosystem composed of "general school support, infrastructure, professional development, teacher attitude, and teacher personal use" (Besnoy et al. 2012). In such an ecosystem, the teachers use technology to aid them to prepare coursework, grade assignments, and communicate with students while students can interact digitally while allowing students to also explore many different technology applications (Palak and Walls 2009).

Computer Science

In computer science, digital ecosystem refers to the two-level system of interconnected machines (Briscoe and Wilde 2009). At the first level, optimization and computing services take place over a decentralized network which feeds into a second level that operates on a local level that seeks to operate within local constraints. Local searches and computations can be performed more efficiently as a result of this process because of the requests first being handled by other peers with similar constraints. This scalable architecture is referred to as a digital ecosystem that builds upon the concept of "service-oriented architecture with distributed evolutionary computing." Different from the other domains' ecosystems, in this model of ecosystem, the actors within the ecosystem are applications or groups of services (Briscoe and Wilde 2009).

Artificial Life and Intelligence

The study of artificial life also has the concept of digital ecosystem which came into fruition in late 1996 with the creation of an artificial life entertainment software product called *Creatures*. In this example, the digital ecosystem was comprised of users who all interact with one another, essentially creating an online persona separate from their own (Cliff and Grand 1999).

Additionally, smart homes, homes in the basic elements such as air conditioning or security systems can be controlled via wireless technology, are considered to be digital ecosystems as well due to their interconnected nature and reliance upon one another to make decisions using artificial intelligence (Harper 2003). In this instance, the actors that make up the digital ecosystem are the independent components that interact with one another supported by a collection of knowledge that then can be disseminated amongst the ecosystem (Reinisch et al. 2010).

The Future Research on Digital Ecosystem

From a general digital ecosystem perspective, future research could focus on adding additional

potential applications for studying digital ecosystems such as the healthcare industry, manufacturing, retail, or even politics, etc., especially as they pertain to big data. As in all contexts defined, digital ecosystems are comprised of highly interconnected people and/or machines and as such their specific relations to one another. Big data focuses on finding overarching trends by bridging together disparate data sources and creating profiles of sorts in different contexts. That said, digital ecosystem research could play a pivotal role in unlocking new avenues to discover new trends which could aid future data science research and companies as well.

Additionally, as digital ecosystems can bridge the gap of space and distance, new research could be conducted to understand digital ecosystems in an international context. Most of studies which were covered did not really consider the impact of the interaction between countries via a digital ecosystem could play on how that ecosystem performs in each country's environment. Therefore, digital ecosystems impact in an international context could help shed light on these digital ecosystems.

Conclusion

Dependent on technology, digital ecosystem connects people and machines. The concept has been applied to various domains including business, education, artificial intelligence, etc. Digital ecosystems provide platforms for big data to be produced and exchanged.

Further Reading

- Besnoy, K. D., Dantzler, J. A., & Siders, J. A. (2012). Creating a digital ecosystem for the gifted education classroom. *Journal of Advanced Academics*, 23(4), 305–325.
- Briscoe, G., & De Wilde, P. (2009) Digital ecosystems: Evolving service-oriented architectures. [arXiv.org](https://arxiv.org/abs/0908.0001).
- Cliff, D., & Grand, S. (1999). The creatures global digital ecosystem. *Artificial Life*, 5(1), 77–93.
- Harper, R. (2003). *Inside the smart home*. Bristol: Springer.
- Li, W., Du, W., & Yin, J. (2017). Digital entrepreneurship ecosystem as a new form of organizing: The case of

Zhongguancun. *Frontiers of Business Research in China*, 11(1), 69–100.

- Palak, D., & Walls, R. T. (2009). Teachers' beliefs and technology practices. *Journal of Research on Technology in Education*, 41(4), 417–441.
- Reinisch, C., Kofler, M. J., & Kastner, W. (2010). ThinkHome: A smart home as digital ecosystem. In *Conference proceedings in 4th IEEE international conference on digital ecosystems and technologies*. Dubai: United Arab Emirates (pp. 12–15). https://books.google.com/books/about/4th_IEEE_International_Conference_on_Dig.html?id=2AgunQAACAAJ.
- Sussan, F., & Acs, Z. (2017). The digital entrepreneurial ecosystem. *Small Business Economics*, 49, 55–73.
- Tansley, A. G. (1935). The use and abuse of vegetational concepts and terms. *Ecology*, 16, 284–307.
- Weill, P., & Woerner, S. L. (2015). Thriving in an increasingly digital ecosystem. *MITSloan Management Review*, 56(4), 27–34.

Digital Inequality

► Digital Divide

Digital Knowledge Network Divide (DKND)

Connie L. McNeely and Laurie A. Schintler
George Mason University, Fairfax, VA, USA

In general, the “digital divide,” as a term, has referred to differential access to digital means and content and, as a phenomenon, increasingly affects the ways in which information is engaged at a most basic level. However, the digital divide is becoming, more fundamentally, a “knowledge divide.” Knowledge implies meaning, appropriation, and participation, such that access to knowledge is a means to achieve social and economic goals (UNESCO 2005). In this sense, the knowledge divide indicates a growing situation of relative deprivation in which, as in other societal domains, some individuals and groups reflect lesser capacities relative to others to access knowledge for social benefit and contribution. Furthermore, today's knowledge society encompasses a system of highly complex and

interconnected networks. These are digital networks marked by growing diversification in information and communication technology (ICT) capacities by which data generation and diffusion translate into differences in overall access and participation in the knowledge society. Related conceptions of this networked knowledge society rest on visions of a world in which ICTs contribute to organizational and social structures by which access and participation are differentially available to various member in society (cf. Castells 2000). To more fully capture the effective dimensions of these differentiating and asymmetric relations, the more explicit notion of the *Digital Knowledge Network Divide* (DKND) has been posited to better describe and understand related structures, dynamics, and relationships (Schintler et al. 2011).

Increasingly characterized by big data derived from social actors and their interactions within and across levels of analysis, DKND culminates in a situation that reflects real world asymmetries among privileges and limitations associated with stratified societal relations. Referring to the explosion in the amounts of data available for research, governance, and decision making, big data is one of the most prominent technology and information trends of the day and, more to the point, is a key engine for social, political, and economic power and relations, creating new challenges and vulnerabilities in the expanding knowledge society. In fact, the collection, analysis, and visualization of massive amounts of data on politics, institutions and culture, the economy, and society more generally have important consequences for issues such as social justice and the well-being of individuals and groups, and also for the stability and prosperity of countries and regions, as found in global North and South (or developed and developing) country divides. Accordingly, a comprehensive understanding of DKND involves consideration of various contexts for and approaches to bridging digital, knowledge, and North-South divides.

Such divides present critical expressions of relative inequalities, inequities, and disparities in information (which may include misinformation and disinformation) and knowledge creation,

access, opportunities, usage, and benefits between and among individuals, groups, and geographic areas. Thus, the DKND must be understood in keeping with its many guises and dimensions, which means considering various perspectives on related capacities to explore broader implications and impacts across global, national, and regional contexts and levels of analysis. Framed relative to capacities to not only access and engage data, but also to transform it into knowledge and actionable insights, DKND is determined by such issues as inequalities in digital literacy and access to education. These issues affect, for example, scientific mobility, smart technologies and automation, labor market structures and relations, and diversity within and across different types and levels of socio-technological engagement and impact. In particular, digital literacy and digital access are fundamental to conceptualizing and understanding the basic dimensions of the DKND.

Digital Literacy

Undergirded by the growth of big data, knowledge intensification and expansion are raising concerns about building digital literacy, especially as a key DKND determinant. Indeed, the “online/not online” and technology “have/have not” focus of many digital divide discussions obscures a larger digital equity problem: disparities in levels of digital readiness, that is, of digital skills and capacities (Horrigan 2019). The knowledge divide also is a skills divide and, while demands are being issued for a more highly educated and digitally literate population and workforce, opportunities for participation and mobility are at the same time highly circumscribed for some groups. In developing knowledge, big data analytics, or the techniques and technologies needed for harnessing value from use of data, increasingly define digital literacy in this regard. On the one hand, digital literacy is about enabling and putting technology to use. However, on the other hand, questions of digital literacy move the issue beyond base access to hardware and technology. Digital literacy is intimately about the capabilities

and skills needed to generate knowledge, to use relevant hardware and technology to help extract information and meaning out of data while coping with its volume, velocity, and variety, and also its variability, veracity, vulnerability, and value (the “7 Vs” of big data).

More than access to the basic digital infrastructure needed to benefit from big data, digital literacy is “the ability to use information and communication technologies to find, evaluate, create, and communicate information, requiring both cognitive and technical skills” (ALA 2020). Access to and use of relevant technologies arguably can provide an array of opportunities for digital literacy and knowledge-enhancing options. Digital literacy is about making meaning and can be considered a key part of the path to knowledge (Reedy and Parker 2018), with emphasis on digital capacities for finding, creating, managing, processing, and disseminating knowledge. In the ideal, a basic requirement for digital literacy is high-quality education for all, and the growth of digital networks, which are at the core of the knowledge society, opening opportunities to facilitate education and learning. However, even if there is technical access to digital information and networks, those features may not be meaningful or commensurate in people’s everyday lives with education and learning opportunities (Mansell and Tremblay 2013), which are the means for broader digital access.

Digital Access

Access is the primary allocating factor determining network entry and participation and can be cast in terms of at least four interrelated issues relative to big data generation and use: 1) physical and technical means; 2) learning and cognitive means; 3) utilization preferences and styles as means to different types of data and knowledge; and 4) the extent and nature of those means. These issues speak to how knowledge networks are engaged, referencing differences in access to data and knowledge, with implications for societal participation and contributions and for benefit as opposed to disadvantage. Differences in data

access, use, and impact are found within and across levels of analysis. Moreover, crossover and networked big data challenge notions of consent and privacy, with affected individuals and groups having little recourse in disputing it. Those with access in this regard have capacities to intervene to mitigate the gaps and disparities linked to big data and, in turn, the information and power imbalances dictated by those with control and are the major users of big data and the intelligence produced from it and related technologies and analytics. This point again leads to questions concerning who has access to the data and, more generally, who benefits.

Related inequalities and biases are reflected in gatekeeping processes reflected in the broader societal context, such that there exists not only a digital divide but, more specifically, knowledge brokers and a DKND in which asymmetry is the defining feature (Schintler et al. 2011). As mentioned, the digital divide generally has been characterized as a critical expression of relative inequalities and differences in ICT access and usage between and among individuals, groups, and geographic areas. However, it more broadly references gaps within and among digital “haves” and “have-nots” and the digitally literate and illiterate in different socio-spatial contexts. In this sense, the DKND is constituted by various modalities of inequality and points to the notion that it encompasses multiple interactive relational structures enacted in diverse institutional frames in which access to education, skills, and ICT capabilities and opportunities are variably distributed throughout social, political, and economic systems.

Some depictions of knowledge societies posit that they ideally should be concerned not only with technological innovation and impacts, but they are defined in concert with human development resting particularly on universal access to information and knowledge, quality education for all, and respect for diversity – that is, on a vision of promoting knowledge societies that are inclusive and equitable (Mansell and Tremblay 2013). In a digitally enabled big data world, many human activities depend on how and the degree to which information and knowledge are

accessed, generated, and processed. Moreover, the capacity, motivation, education, and quality of knowledge acquired online have consequences for life opportunities in the social realm (Ragnedda 2019). Different capacities for digital access and use can translate to different roles of big data for individuals, communities, and countries, strongly influencing inequalities and related divides. Also, open, re-used, and re-combined data can bring both opportunities and challenges for society relative to social, economic, and international dimensions, with implications for equity and social cohesion. Big data pervasiveness is linked to the rise and persistence of digital divides and inequalities that reflect impacts of and on structural relations determining the access and engagement of related knowledge and network resources.

In keeping with varying capacities and possibilities to transform digitally valuable resources and knowledge into social and tangible benefits (Ragnedda 2019), different access and different abilities and skills for exploiting ICT-related benefits are strongly connected with societal inequalities understood relative to physical, financial, cognitive, production, design, content, institutional, social, and political access – all of which can operate to create or reinforce divides in digital experiences and related outcomes (Ragnedda 2019; DiMaggio et al. 2004). Although this situation need not be framed as static reproduction, understanding societal dynamics means that, despite a relatively open internet, everyone is not in the same position to access and use opportunities offered in the digital arena (Ragnedda 2019). Even with better skills and qualifications – the acquisition of which is affected by social dynamics and structures – previous societal positions and relationships influence capacities to access related opportunities in the social realm.

Conceptual Scope

Not only technological but also social, economic, and political developments mark the parameters

of the DKND. These aspects are highly interrelated, operating interdependently relative to various divides in regard to impact on society and the world. Over the last several years, digital content has been growing at an astronomical rate and it is in this sense that big data diversification and complexity feeds into the creation and diffusion of knowledge as a networked process. User-generated data networks, and the types of information or knowledge to which they may or may not have access, constitute complex digital divides. Networks are indeed an integral feature of the knowledge divide, constituting complex systems in which some individuals or groups are more central or have more influence and control over the creation and flow of knowledge and information. The flow and manipulation of data and information to create knowledge are dynamic processes, influencing access to and participation in digital knowledge networks. Participation in these digital networks can have variable effects. For example, the use of algorithms that determine information access based on past activities can result in filter bubbles that cause limited access and intellectual isolation (Pariser 2012). Another example is akin to the “Matthew effect,” in which “the rich get richer and the poor get poorer,” referencing cumulative advantage and gaps between haves and have-nots (Merton 1988). This situation reflects inequalities embedded in knowledge networks (epistemic networks) where the status of some members relative to others is elevated, contributing to inequalities in terms of digital capabilities to access knowledge and to disseminate and receive recognition for it (Schintler and McNeely 2012).

As discussed, networks can offer opportunities for empowerment of marginalized and excluded groups, but possibilities for those opportunities must be understood relative to discrimination, privacy, and ethical issues. With increasing digitization of everything from retail and services to cities and healthcare, and the growth of the Internet of Things (IoT), the emergence of network haves and have-nots is more specifically tied to current digital divides. Networks, as collaborative structures, are central to stimulating the produc-

tion of knowledge beneficially relevant for those who can access and apply it; they can offer opportunities or can block knowledge sharing. The knowledge divide represents inequalities and gaps in knowledge among individuals and groups. The digital divide extends this idea, distinguishing among those with and those without access to the internet. Accordingly, the DKND is conceived via structural relations and dynamics that look beyond technological applications to consider institutional, regulatory, economic, political, and social conditions that frame the generation of digital, knowledge, and network relationships. That is, the DKND reflects networked disparities and gaps in knowledge and associated value that also operate to differentially restrict or enhance access and participation of certain segments of the population.

However, understanding the DKND also requires a state-of-the-art perspective on other aspects of digital networks, pointing to how they appear today and can be expected to do so increasingly in the future. Reference here is to humans interacting with humans mediated by machines (social machines) and, importantly, machines interacting with machines. Indeed, the IoT is all about machine-to-machine (M2M) interactions, exchanging data and information and producing knowledge by advanced computational modeling and deep learning. Note that there is a profound global divide in terms of who has access to the machine hardware and software needed to plug into the IoT. Moreover, there are information/knowledge asymmetries between machines and humans, and even machines and machines (Schintler 2017).

Conclusion

Inequality marks big data, reflected in hierarchically differentiated structures defined and privileged according to those who create the data, those who have the means to collect it, and those who have the expertise to analyze and use it (Manovich 2011; Schintler et al. 2011). The extent

to which massive amounts of data translate into information, and that into expanded knowledge, speaks to ways in which big data is being used and framed as sources of discovery and knowledge creation. Big data, as a broad domain of practice and application, can be understood relative to advancing knowledge and productivity (Schintler and McNeely 2012). It is in this regard that networks reflect defining processes that address both formal and informal relationships among different actors ranging from individuals to countries. This situation does not occur in isolation and, as such, necessitates a comprehensive view on the promises and challenges attached to big data and the diffusion of knowledge and establishment of networks determining digital relations.

The DKND reflects a complex and adaptive system bound by socio-technological structures and dynamics that largely depend on access, cognitively, normatively, and physically determined across different levels and units of analysis. Critical dimensions of this perspective include relational and territorial digital knowledge network formation, characteristics, and effects; digital knowledge network opportunity structures and differentiation; and overall vertical and horizontal trends and patterns in the DKND. As such, it has broad implications for ways of thinking about data, their sources, uses, and purposes. The DKND brings particular attention to the extent to which big data might exacerbate already rampant disparities, pointing to how big data are used by different actors, for what purposes, and with what effects. In fact, this is the big data divide and related developments have been at the center of critical public and intellectual debates and controversies.

By definition, the DKND operates in accordance with the social contexts in which big data analytics are engaged and applied, and these relations can be considered in terms of institutional frameworks, governance conditions, and system dynamics. Big data analytics are enabled by technical advances in data storage capacities, computational speeds, and the near real-time availability of massive datasets. The ability to integrate and

analyze datasets from disparate sources and to generate new kinds of knowledge can be beneficial, but also can constitute legal, ethical, and social dilemmas leading to hierarchical asymmetries. Considered relative to questions of social and structural dynamics and material capacities, different perspectives on digital asymmetries and related effects can be framed in terms of the evolving socio-technological landscape and of disparities grounded in broader societal and historical dynamics, relationships, and structures that constrain equal and equitable outcomes.

More to the point, big data has led to profound changes in the way that knowledge is generated and utilized, underlying the increasingly deep penetration and systems nature of related developments in human activities. Accordingly, the idealized vision of the knowledge society is one in which the full potential of digital networks is achieved in an equitable and balanced knowledge environment – one in which knowledge is integrated in ways that maximize benefits and minimize harms, taking into account goals of social, economic, and environmental wellbeing (Mansell and Tremblay 2013). However, broad socioeconomic characteristics are basic factors affecting capacities for realizing digital literacy, access, and engagement, differentially positioning and enabling individuals, groups, and countries to capture knowledge benefits.

Big data capacities and possibilities for digital access, broadly defined, are affected by the DKND, which determines and is determined by the character, types, and consequences of differentiated digital access and opportunities. The digital divide in general is a multifaceted phenomenon, interwoven with existing processes of social differentiation and, in fact, may accentuate existing inequalities (Ragnedda 2019). While, at a fundamental level, the digital divide is based on technological and physical access to the internet and related hardware, knowledge and networks further affect participation and consequences also tied to already existing social inequalities and gaps. In the face of digital, knowledge, and network asymmetries, the DKND stands in contrast to broader visions of societal equity and wellbeing, reflecting sensitivity to the complexities

and realities of life in the big data knowledge society.

Further Reading

- American Library Association (ALA). (2020). Digital literacy. <https://literacy.ala.org/digital-literacy>.
- Castells, M. (2000). *Rise of the network society*. Malden: Blackwell.
- DiMaggio, P., Hargittai, E., Celeste, C., & Shafer, S. (2004). Digital inequality, from unequal access to differentiated use. In K. Neckerman (Ed.), *Social inequality* (pp. 355–400). New York: Russell Sage Foundation.
- Horrigan, J. B. (2019, August 14). Analysis: Digital divide isn't just a rural problem. *Daily Yonder*. <https://dailyyonder.com/analysis-digital-divide-isnt-just-a-rural-problem/2019/08/14>.
- Manovich, L. (2011). Trending: The promises and the challenges of big social data. <http://manovich.net/index.php/projects/trending-the-promises-and-the-challenges-of-big-social-data>.
- Mansell, R., & Tremblay, G. (2013). *Renewing the knowledge societies vision: Towards knowledge societies for peace and sustainable development*. Paris: UNESCO. <http://eprints.lse.ac.uk/id/eprint/48981>.
- Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis*, 79, 606–623.
- Pariser, E. (2012). *The filter bubble: How the new personalized web is changing what we read and how we think*. New York: Penguin.
- Ragnedda, M. (2019). Reconceptualizing the digital divide. In B. Mutsaers & M. Ragnedda (Eds.), *Mapping the digital divide in Africa: A mediated analysis* (pp. 27–43). Amsterdam: Amsterdam University Press.
- Reedy, K., & Parker, J. (Eds.). (2018). *Digital literacy unpacked*. Cambridge, UK: Facet. <https://doi.org/10.29085/9781783301997>.
- Sagasti, A. (2013). The knowledge explosion and the knowledge divide. <http://hdr.undp.org/sites/default/files/sagasti-1-1.pdf>.
- Schintler, L. A. (2017). The constantly shifting face of the digital divide. *Big Data for Regional Science*, 28, 336.
- Schintler, L., & McNeely, C. L. (2012). Gendered science in the 21st century: The productivity puzzle 2.0? *International Journal of Gender, Science and Technology*, 4 (1), 123–128.
- Schintler, L., McNeely, C. L., & Kulkarni, R. (2011). Hierarchical knowledge relations and dynamics in the “Tower of Babel.” In *Rebuilding the mosaic: Fostering research in the social, behavioral, and economic sciences at the National Science Foundation in the next decade (SBE 2020)*, NSF 11–086. Arlington: National Science Foundation. http://www.nsf.gov/sbe/sbe_2020.
- United Nations Educational, Scientific, and Cultural Organization (UNESCO). (2005). *UNESCO world report: Towards knowledge societies*. Paris: UNESCO.

Digital Literacy

Dimitra Dimitrakopoulou
School of Journalism and Mass Communication,
Aristotle University of Thessaloniki,
Thessaloniki, Greece

Digital literacy means having the knowledge and the skills to use a wide range of technological tools in order to read and interpret various media messages across different digital platforms. Digital literate people possess critical thinking skills and are able to use technology in a strategic way to search, locate, filter, and evaluate information; to connect and collaborate with others in online communities and social networks; and to produce and share original content on social media platforms. In the era of big data, digital literacy becomes extremely important as internet users need to be able to identify when and where personal data is being passively collected on their actions and interactions and form patterns on their online behavior, as well as contemplate the ethical dilemmas on data-driven decisions for both individuals and society as a whole.

The interactive platforms that the web has introduced to the fields of communication, content producing and sharing as well as to networking, offer great opportunities for the learning and educational procedure for both educators and students. The expanding literature on the “the Facebook generation” indicates a global trend in the incorporation of social networking tools for connectivity and collaboration purposes among educators, students, and between these two groups. The use of social software tools holds particular promise for the creation of learning settings that can interest and motivate learners and support their engagement, while at the same time addressing the social elements of effective learning. At the same time, it is widely suggested that today’s students require a whole new set of literacy skills in the twenty-first century.

The current generation learners, namely, young people born after 1982, have been and are being raised in an environment that presupposes that

new technologies are a usual part of their daily lives. For them the Internet is part of the pattern of their day and integrated into their sense of place and time.

Social web presents new possibilities as well as challenges. On the one hand, the main risks of using the Internet can be classified to four levels: (a) commercial interests, (b) aggression, (c) sexuality, and (d) values/ideology. On the other hand, the web opens a whole new world of opportunities for education and learning, participation and civic engagement, creativity, as well as identity and social connection. Wikis, Weblogs, and other social web tools and platforms raise possibilities for project-based learning and facilitate collaborative learning and participation among students and educators. Moreover, project-based learning offers many advantages and enhances skills and competencies.

The changes in the access and management of information as well as in possibilities for interactivity, interaction, and networking signal a new learning paradigm that is created due to the need to select and manage information from a vast variety of available sources, while at the same time learning in the digital era is collaborative in nature and the learner is no more a passive recipient of information but as an active author, co-creator, evaluator, and critical commentator.

The abovementioned changes signify the foundations for Learning 2.0, resulting from the combination of the use of social computing to directly enhance learning processes and outcomes with its networking potential. The changes that we are experiencing through the development and innovation that the interactive web introduces are framed by the participatory culture that we live in.

Participatory culture requires new literacies that involve social skills which are developed through collaboration and networking. In this environment with new opportunities and new challenges, it is inevitable that new skills are also required, namely, play, performance, simulation, appropriation, multitasking, distributed cognition, collective intelligence, judgment, transmedia navigation, networking, and negotiation.

Nevertheless, the participatory culture is prospectively participatory for all, e.g., providing and enabling all to an open and equal access as well as to a democratized and regulated environment. Three core problems are identified as the main concerns in the digital era, which are addressed by Jenkins et al.:

- (a) *Participation gap*: Fundamental inequalities in young people's access to new media technologies and the opportunities for participation they represent
- (b) *Transparency problem*: Children are not necessarily reflecting actively on their media experiences and cannot always articulate what they learn from their participation.
- (c) *Ethics challenge*: Children cannot develop on their own the ethical norms needed to cope with a complex and diverse social environment online (Jenkins et al. 2006: 12, see more on pp. 12–18).

The necessity to deal with these challenges calls for a twenty-first century media literacy, which can be described as the set of abilities and skills where aural, visual, and digital literacy overlap. These include, as the New Media Consortium indicates, the ability to understand the power of images and sounds, to recognize and use that power, to manipulate and transform digital media, to distribute them pervasively, and to easily adapt them to new forms.

Pupils that still attend school are growing in a technology dominated world. Youth born after 1990 are currently the largest generation in the last 50 years and live in a technology saturated world with tools such as mobile phones and instant access to information. Moreover, they have become avid adopters of Web 2.0 and beyond technologies such as podcasting, social networking, instant messaging, mobile video/gaming, IPTV. Being the first generation to grow up surrounded by digital media, their expectations of connectivity are high, with technology everywhere in their daily life. The characteristics of the new generation of students include, among others, multi-tasking, information age mindset, eagerness for connectivity, “fast-track” accomplishments, preference towards doing than knowing, approach

of “reality” as no longer real, blur lines between the consumer and the creator and expectations for ubiquitous access to the Internet.

These characteristics should be definitely taken into account when designing or evaluating a digital literacy program. Children use the Internet mainly as an educational resource; for entertainment, games, and fun; information seeking; and social networking and share experiences with others. Communication with friends and peers, especially, is a key activity. They use different tools such as chats, instant messaging, or e-mail to stay in contact with each other or to search for new friends. They also participate in discussion forums or use the Internet to search for information, to download music or videos, and to play online games. Communication and staying in touch with friends and colleagues is ranked highly for them.

Learning 2.0 is an emergent phenomenon, fostered by bottom-up approach take up of Web 2.0 in educational contexts. Although social computing originated outside educational institutions, it has huge potential in formal Education and Training (E&T) for enhancing learning processes and outcomes and supporting the modernization of European E&T institutions. Learning 2.0 approaches promote the technological, pedagogical, and organizational innovation in formal Education and Training schemes. As Redecker indicate, the interactive web builds up the prospects for (a) enhancing innovation and creativity, (b) improving the quality and efficiency of provision and outcomes, (c) making lifelong learning and learner mobility a reality, and (d) promoting equity and active citizenship.

On the other hand, there are major challenges that should be dealt with. While there are currently vast numbers of experimental Learning 2.0 projects under way all over the world, on the whole, Learning 2.0 has not entered formal education yet. The following technical, pedagogical, and organizational bottlenecks have been identified by Redecker et al. which may hinder the full deployment of Learning 2.0 in E&T institutions in Europe: (a) access to ICT and basic digital skills, (b) advanced digital competence, (c) special needs, (d) pedagogical skills, (e) uncertainty, (f) safety and privacy concerns, and (g) requirements on institutional change.

Cross-References

- Curriculum, Higher Education, Humanities
- Digital Knowledge Network Divide (DKND)
- Education and Training
- Information Society

Further Reading

- Hasebrink, U., Livingstone, S., & Haddon, L. (2008). *Comparing children's online opportunities and risks across Europe: Cross-national comparisons for EU Kids Online*. Deliverable D3.2. EU Kids Online, London. Retrieved from http://eprints.lse.ac.uk/21656/1/D3.2_Report-Cross_national_comparisons.pdf.
- Jenkins, H., et al. (2006). *Confronting the challenges of participatory culture: Media education for the 21st century*. Chicago: The MacArthur Foundation. Retrieved from http://digitallearning.macfound.org/atf/cf/%7B7E45C7E0-A3E0-4B89-AC9C-E807E1B0AE4E%7D/JENKINS_WHITE_PAPER.PDF.
- Literacy Summit. (2005). *NMC: The New Media Consortium*. Retrieved from http://www.nmc.org/pdf/Global_Impervative.pdf.
- Redecker, C., et al. (2009). *Learning 2.0: The impact of web 2.0 innovations on education and training in Europe*. Final Report. European Commission: Joint Research Centre & Institute for Prospective Technological Studies. Luxembourg: Office for Official Publications of the European Communities. Retrieved from <http://ftp.jrc.es/EURdoc/JRC55629.pdf>.

Digital Storytelling, Big Data Storytelling

Magdalena Bielenia-Grajewska
Division of Maritime Economy, Department of
Maritime Transport and Seaborne Trade,
University of Gdansk, Gdansk, Poland
Intercultural Communication and
Neurolinguistics Laboratory, Department of
Translation Studies, University of Gdansk,
Gdansk, Poland

Storytelling dates back to the ancient times when stories were told among the members of communities, using oral, pictorial, and, later, also writing systems. The reminiscent of the first examples of storytelling from the past centuries can still be

observed on the walls of buildings coming from the antiquity or on the parchments originating from the dim and distant times. Nowadays, people also tell stories in private and professional life. Shell and Moussa (2007) stress that there are certain aspects that make stories interesting and effective. What differentiates a story from an example is dynamism. When one listens to the story, he/she starts to follow the plot and think what happens next. Allan et al. (2001) state that stories stimulate imagination. The narrative approach points to the items that are often neglected, such as the ones that are simpler or less precise. As Denning claims, “a knowledge-sharing story describes the setting in enough detail that the solution is linked to the problem by the best available explanation” (2004: 91).

The important caesura for storytelling was the invention of the print that facilitated the distribution of narrations among a relatively large group of people. The next crucial stage was the rapid development in the technological sphere, represented by, e.g., the introduction and proliferation of online technologies. Digital Storytelling can be defined as the application of technological advancements in telling stories, being visible in the usage of computer-related technologies in documenting events or narrating one's personal experience. Pioneers in Digital Storytelling included the following people: Joe Lambert, who co-founded the Center for Digital Storytelling (CDS) in Berkeley and Daniel Meadows who is the British photographer, author and a specialist in education. Also, alternative names for this phenomenon include *digital documentaries*, *computer-based narratives*, *digital essays*, *electronic memoirs* or *interactive storytelling*. The plethora of names shows how different functions digital storytelling may have; it may be used to document a story, narrate an event or act as a diary in the computer-related reality.

Moreover, technology has influenced storytelling also in another way: by creating and distributing big data. Thus, digital storytelling focuses not only on presenting a story but also on displaying big data in an efficient way. However, it should be mentioned that since this method of data creation and dissemination can be used by individuals regardless of their knowledge, technical

capabilities and attentiveness to proper online expression, digital storytelling offered in the open access mode can be of different quality. The most common failures that can be observed as far as the production of digital storytelling is concerned are the lack of presentation skills and technical abilities to create an effective piece of storytelling. Another factor that may influence the perception and cognition of storytelling is the absence of adequate linguistic skills. Digital Storytelling created by a person who makes language mistakes and has incomprehensible pronunciation is not likely to have many followers and its educational application is also limited. In addition, the effective presentation of big data in digital storytelling is a difficult task for individuals not familiar with big data analytics and management. There are certain elements that are connected with a success or failure of this mode of expression. At the website called “Educational Uses of Digital Storytelling,” seven important elements of digital storytelling are mentioned. The first one is *Point of View* and it is connected with presenting the main point of the story and the approach adopted by the author. The second issue – *A Dramatic Question* – concerns the main question of the story to be answered in the last part of storytelling. The third notion, *Emotional Content*, reflects the emotional and personal way of presenting a story that makes target viewers involved in the plot. The fourth element named *The Gift of Your Voice* encompasses the strategies aimed at personalizing the storytelling that facilitates the understanding of the story. The fifth notion is connected with the audio dimension of digital storytelling; *The Power of the Soundtrack* is related to the usage of songs, rhythms, and jingles to make the story more interesting and informative. The next one called *Economy* is related to the amount of material presented for the user; digital storytelling should not bore viewers because of its length and the immense amount of content. The last element, *Pacing*, is connected with adjusting the rhythm to the presented content.

It also should be mentioned that a good example of digital storytelling should offer a combination of audio, pictorial, and verbal representation that is coherent and adopted for the target

audience. As far as the requirements and possibilities of the viewers are concerned, the content itself and the method of presenting the content should be fitting to their age, linguistic skills, education, and occupation. For example, a piece of digital storytelling recorded to master English should be produced by taking into account the age of learners, their level of linguistic competence as well as professional background (especially in the case of teaching English for Specific Purposes). Another important feature of digital storytelling in relation to the immensity of information that should be covered in the film lasting from 2 to 10 min is linked with using and presenting big data in an efficient way. Apart from storytelling, there are also other names used to denote telling stories in different settings. Looking at the organizational environment, Henderson and Boje (2016) discuss the phenomenon of fractal patterns in quantum storytelling and Big Story by Mike Bonifer and his colleagues.

Tools in Digital Storytelling

Tools used in digital storytelling can be classified into verbal and nonverbal ones. Verbal tools encompass all types of linguistic representation used in telling stories. They include the selective use of words, phrases, and sentences to make the piece of information more interesting for the target audience. The linguistic dimension can be further investigated by applying the micro, meso, or macro perspectives. The micro approach is connected with analyzing the role of a single word in the processes of creation and cognition of information. For example, adjectives are very powerful in creating an image of a thing or a person. Such adjectives as *prestigious*, *unique*, or *reliable* stress the high quality and effectiveness of a given offer. When repeated throughout the piece of digital storytelling, they strengthen the identity of a company offering such products or services. Numerals constitute another effective tool of digital storytelling. The same number presented in different numerical representation may have a different effect on the target audience. For example, the dangerousness connected with the

rising death toll of a contagious disease is perceived in a different way when presented with percentage (e.g., 0.06% infected) and when described by numbers (e.g., 100,000 infected). Another example may include organizational discourse and the usage of numerals to stress the number of customers served every month, the amount of early income, etc. In the mentioned case, big numbers are used to create the image of a company as a leader in its industry, being a reliable and an efficient player on the market. The meso dimension of investigating focuses on structures used to decode and encode messages in digital storytelling. Taking the grammar perspective into account, active voice is used to stress the personal involvement of speakers or the described individuals in the presented topic (e.g., *we have made* instead of *it was made*). Active voice is often used to stress responsibility and devotion to the discussed phenomenon. Moreover, questions are used in digital storytelling to draw the viewers' attention to the topic. The macro approach, on the other hand, is related to the selection of texts used in digital storytelling. They include, among others, stories, interviews, descriptions, presentations of websites, and other online textual forms.

Verbal tools can also be subcategorized into literal and nonliteral methods of linguistic expression. Literal tools encompass the types of meanings that can be deduced directly, whereas nonliteral communication makes use of speakers' intuition and knowledge. Nonliteral (or figurative) discourse relies, e.g., on metaphors in presenting information. Applying the micro perspective, metaphorical names are used to tell a story. Relying on a well-known domain in presenting a novel concept turns out to have better explanatory characteristics than the literal way of describing a phenomenon. Apart from the informational function, metaphors, having often some idea of mystery imbedded in them, attract the viewers' attention more than literal expressions. Taking into account the sphere of mergers and acquisitions, such names as *white knight*, *lobster trap*, or *poison pill* describe the complicated strategies of takeovers in just few words. In the case of specialized data, metaphors offer the explanation for

laymen and facilitate the communication between specialists representing different domains. Using the macro approach, metaphors are used to create organizational identity. For example, such metaphors as *organization as a teacher* or *organization as a family* can be constructed after analyzing the content presented in computer-related sources. Discussing the individual level of metaphorical storytelling, metaphors can be used to create the identity of speakers. Forming new metaphors makes the creator of digital storytelling characteristic and outstanding. At the same time, it should be noticed that digital storytelling may facilitate the creation of novel symbolic representations. Since metaphors origin when people observe the reality, digital storytelling, as other types of texts and communication channels, may be a potential source of new metaphors. Thus, digital storytelling may stimulate one's creativity in terms of writing, speaking, or painting.

The sphere of nonverbal tools encompasses mainly the auditory and pictorial way of communicating information. As far as the auditory dimension is concerned, such issues as soundtracks, jingles, sounds as well as the voice of the speaker are taken into account. The pictorial dimension is represented by the use of any type of picture-related representation, such as drawings and pictures. It should be stated that all the mentioned tools should not be discussed in isolation; the power of digital storytelling lies in the effective combination of different representations. It should be stressed that the presentation of audio, verbal and pictorial representations often relies on advanced technology.

The tools used in digital storytelling may also be studied by taking into account the stage of creating and disseminating digital storytelling. For example, the preparation stage may include different qualitative and quantitative methods of data gathering in order to construct a viable representation of the discussed topic. The stage of creating digital storytelling encompasses the application of computer-based technologies to incorporate verbal, audio and pictorial representations into storytelling. The discussion on tools used in digital storytelling should also encompass the online and offline methods of disseminating

the produced material. As far as the online dimension is concerned, it includes mainly the application of social networking tools, discussion forums, websites and newsletters that may provide the piece of digital storytelling itself or the link to it. The offline channels, such as books, newspaper articles or corporate leaflets include mainly the link to the piece of digital storytelling. As Ryan (2018) mentions, since modern storytelling focuses much on numbers, we can talk about visual data storytelling.

Methodologies of Studying Digital Storytelling

The investigation on digital storytelling should focus on all dimensions related to this form of communication, namely, the verbal, audio, and pictorial ones. Since digital storytelling relies on pictures, the method called video ethnography, used to record people in their natural settings, facilitates the understanding how films are made. To research the verbal sphere of digital storytelling, one of the approaches used in text studies may be applied. These include ethnographic studies, narrative analysis, narrative semiotics, or functional pragmatics. At the same time, an attempt should be made to focus on an approach that may investigate at least two dimensions simultaneously as well as the interrelation between them.

One of the methods to research digital storytelling from more than one perspective is to adopt the Critical Discourse Analysis that focuses not only on the verbal layer of a studied text but also on the relation between nonverbal and verbal representation and how their coexistence determines the way a given text is perceived. This approach provides information for the authors of digital storytelling how to make their works reach a relatively large group of people. The complexity and multifactorial character of digital storytelling can be researched by using network theories. For example, Social Network Analysis studies the relations between individuals. In the case of digital storytelling, it may be applied to investigate the relations between the individuals speaking in

the piece of digital storytelling as well as between the interlocutors and the audience.

Another network approach, Actor-Network-Theory, is used to stress the importance of living and non-living entities in the creation and performance of a given phenomenon. In the case of digital storytelling, it may be used to study the role of technological advancements and human creativity in designing a piece of digital storytelling. Since digital storytelling is aimed at creating some emotions and response among the target audience, both researchers and creators of digital storytelling are interested what linguistic, audio and pictorial tools make the piece of digital storytelling more effective. In the mentioned case, the researchers representing cognitive studies and neuroscience may offer an in-depth analysis of constituting elements. For example, by observing the brain or the nervous systems scientists may check the reactions of individuals to the presented stimuli. Using such neuroscientific equipment as magnetoencephalography (MEG), transcranial magnetic stimulation (TMS) functional magnetic resonance imaging (fMRI), eye-tracking or galvanic skin response offers the studies of ones' reactions without the fear of facing fake answers that they may sometimes happen in standard interviews or surveys.

Applications and Functions of Digital Storytelling

The functions of digital storytelling can be divided into individual and social ones. The individual role of digital storytelling is connected with informing others about one's personal issues as well as giving vent for emotions, opinions and feelings. It is often used by people to master their own speaking or presentation skills as well as to exercise technical and computer passion. The social dimension of digital storytelling encompasses the role of digital storytelling in serving functions other than the purely personal ones. For example, digital storytelling is used in education. It draws the attention of students to important issues, by using diversified (e.g., verbal and non-verbal) ways to tell a story. This method of

teaching may also result in the active participation of participants; students are not only the passive recipients of displayed issues, but they are also capable of constructing their own stories. As far as the publication outlets are concerned, there are online platforms devoted to the presentation of digital stories that can be used to enhance one's knowledge on the application of digital storytelling.

The usage of digital storytelling in education can be understood in two ways. One approach is to use digital storytelling to inform viewers about new issues and concepts. Explaining a novel product or a technologically difficult matter by using digital storytelling proves to be more efficient than standard methods of providing information. The second function of digital storytelling is more socially-oriented; digital storytelling may facilitate the understanding of intercultural differences or social issues and make people more sensitive to other people's needs, expectations and problems. It should also be stated that digital storytelling facilitates offering education to those who cannot access the same type of educational package in offline environments. For example, digital storytelling offers education to handicapped people who cannot frequent regular schooling as well as to those who because of geographical or economic distance cannot participate in the regular class. Moreover, digital storytelling may provide knowledge for those who are not interested in participating in standard courses but they want to learn just for pleasure. An example of courses that meet different needs of users is the idea of MOOCs, massive open online courses, being offered on the Internet in the open access mode. Often created by top universities and supported by international organizations, MOOCs are often accessible for free or for a charge if an individual is interested in gaining a proved statement of taking the course, a certificate or ECTS points. By publishing courses on specialized online platforms, MOOCs reach diversified users in different geographical locations who can study the presented content at their own pace.

Another application of digital storytelling concerns the sphere of marketing. Digital storytelling is used by companies to create their identity,

promote their products and services as well as communicate with the broadly understood stakeholders. It should also be mentioned that digital storytelling, being a complex tool itself, may combine different functions at the same time. For example, corporate materials presented by using digital storytelling may serve both marketing and educational functions. The case is the policy of adopting Corporate Social Responsibility that stresses the active involvement of companies in creating and sustaining the harmony with the broadly understood environment. Presenting CSR policies in digital storytelling does not only create the positive image of a company, being an active and supportive member of the community, but also shows how the viewers may take care about the environment themselves. Another social function of digital storytelling is the formation of communities. Those who use digital storytelling may comment the content and express opinions on the topics presented in the recording. Thus, the application of digital storytelling serves many functions for both individuals and organizations.

Big Data Storytelling

Digital storytelling undergoes constant changes and finds new applications due to the rapid development in the sphere of technology. One of the most visible changes can be observed in the sphere of data, being represented by the large and complex datasets as well as handling them (creating, storing, applying, using, updating). Big data may have different forms, such as written, visual, audio, video ones or have more than one form at the same time. Moreover, modern technology allows for changing the form of data into a different one, meeting the needs and expectations of the target audience. The application of big data is connected with a given area of life and type of profession. For example, demographic information, data on business transactions (e.g., purchases) and data on using mobile technology are studied in marketing. The attitude to data nowadays is also different from the one that could be observed in the past. Nowadays companies do not only accumulate information after sales but also

monitor and gather data during operations. For example, logistics companies track the route of their products to optimize services and reduce transportation costs. The attitude to data is also connected with the profile of business entities. For example, companies offering online services deal with data on everyday basis, by managing users' data and monitoring the interest in the offered merchandise, etc. The main functions of gathering and researching big data encompass the opportunity to profile customers and their needs, analyze how often and what they purchase as well as estimate general business trends.

After information is gathered and stored, specialists must take care of presenting it to the audience. Data can be analyzed and models can be created by the use of such programs as, e.g., MATLAB. This program offers signal, image and video processing and data visualizing for engineers and scientists. The visualization of big data may be supported by tools (e.g., Google Maps API) that offer maps and data layers. Such tools provide the visual presentation of such data as, among others, geographical and geospatial data, traffic conditions, public transport data and weather forecasts. These tools facilitate the creation and cognition of big data used in digital storytelling by making immense information compact and comprehensible. Madhavan et al. (2012) discuss the Google Fusion Tables (GFT) offering collaborative data management in the cloud. This type of tool offers reusing existing data gathered from different producers and applying it in different contexts. It is used by, e.g., journalists to visualize some aspects presented in their articles. Maps created by using GFT can be found by readers of various periodicals (e.g., *UK Guardian*, *Los Angeles Times*, *Chicago Tribune*, and *Texas Tribune*).

Further Reading

- Allan, J., Gerard, F., & Barbara, H. (2001). *The power of Tale. Using narratives for organisational success*. Chichester: Wiley.
- Bielenia-Grajewska, M. (2014a). The role of figurative language in knowledge management. Knowledge encoding and decoding from the metaphorical perspective. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology*. Hershey: IGI Publishing.
- Bielenia-Grajewska, M. (2014b). CSR Online Communication: the metaphorical dimension of CSR discourse in the food industry. In R. Tench, W. Sun, & B. Jones (Eds.), *Communicating corporate social responsibility: perspectives and practice (critical studies on corporate responsibility, governance and sustainability)* (Vol. 6). Bingley: Emerald Group Publishing Limited.
- Bielenia-Grajewska, M. (2014c). Corporate online social networks and company identity. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of social network analysis and mining*. Berlin: Springer.
- Denning, S. (2004). *Squirrel Inc. A fable of leadership through storytelling*. San Francisco: Jossey-Bass.
- Henderson, T., & Boje, D. M. (2016). *Organizational development and change theory: managing fractal organizing processes*. Abingdon: Routledge.
- Madhavan, J., et al. (2012). Big data storytelling through interactive maps. *IEEE Data Eng Bull*, 35(2), 46–54.
- Ryan, L. (2018). *Visual data storytelling with tableau: story points, telling compelling data narratives*. Boston: Addison-Wesley Professional.
- Shell, R., & Moussa, M. (2007). *The art of Woo: using strategic persuasion to sell your ideas*. London: Penguin Books Ltd..

Websites

- Educational Uses of Digital Storytelling, <http://digitalstorytelling.coe.uh.edu/>. Accessed 10 Nov 2014.
- Google Fusion Tables., <https://developers.google.com/fusiontables/>. Accessed 10 Nov 2014.
- Matlab. <http://www.mathworks.com/products/matlab/>.

Cross-References

- Content Management System (CMS)
- Digital Literacy
- Humanities (Digital Humanities)
- Knowledge Management
- Online Identity

DIKW Pyramid

- Data-Information-Knowledge-Action Model

Disaster Management

- Natural Hazards

Disaster Planning

Carolynne Hultquist
Geoinformatics and Earth Observation
Laboratory, Department of Geography and
Institute for CyberScience, The Pennsylvania
State University, University Park, PA, USA

Definition/Introduction

Disaster planning is important for all stages in the disaster management cycle, and it occurs at many levels from individuals to communities and governments. Planning for disasters at a large scale requires information of physical and human attributes which can be derived from data collection and analysis of specific areas. This general collection of spatial data can have a large volume and be from a variety of sources. Hazards stemming from different types of natural, man-made, and technological processes require unique planning considerations but can often use a common basis of information to understand the location. A common structure of organization can be adopted despite needing to plan for providing varying resources and implementing procedures required during unique disaster events.

Planning for Disasters

Disaster planning is a crucial part of the disaster management cycle as the planning stage supports all the stages in the process. A hazard is the event itself, such as an earthquake or hurricane, but a disaster is when there is loss of human life and livelihood. Part of the planning is preparation to be less vulnerable to hazards in order for an event not to be a devastating disaster. Information on who is located where and how vulnerable that area is to certain hazards is important to plan for a disaster. It involves considering data on physical processes and human interests and developing disaster response plans, procedures, and processes to make decisions on how to respond during the event and to effectively recover. Planning for a

disaster occurs at many levels from individuals to national and international levels.

Preparation for disasters can be initiated from the ground up as grassroots movements and from the top down as government policy. Community and individual planning normally stems from personal initiatives to initiate a procedure and store basic human needs such as food, water, and medical supplies as disasters can disrupt access to these essentials. Government planning occurs at federal, state, and local levels with the goals of integrating disaster preparation and responses at each level to efficiently use resources and not duplicate efforts. Long-term government planning for hazards can include steps such as analyzing data in order to make strategic decisions that mitigate the hazard and to not contribute to costlier disasters. Disaster planning should include data that are location specific and which will provide relevant information to responders. Geographical Information Systems (GIS) can be used to store data as layers of features representing physical and social attributes. Data collection and analysis does not just start when the event occurs but should be used to help with prediction of risks and assessments of impact. During the event, data being received in real-time can help direct efforts and be fused with pre-existing data to provide context.

One of the goals of planning for disasters is to take steps to be more resilient to hazards. The majority of the time invested in planning for hazards are for those that are likely to occur in the specific area in which people live or that would be the most devastating. Hazards are spatial processes as physical and metrological conditions occur in relation to the features and properties of the Earth. Disasters are inherently linked to human location and involve human-environment interactions; it is clear that a major earthquake and tsunami in Alaska will likely have much less loss of life than if it occurs off the coast of Japan. A spatial relationship exists for both the occurrence of a hazardous event and for the event to have a human impact which makes it a disaster. Impact is primarily measured in human loss but other considerations such as financial and environmental losses are often considered

(Hultquist et al. 2015). Geospatial data are often used to recognize and analyze changes to better understand this relationship through the use of technologies. Monitoring networks are often put in place to identify that events are taking place in order to have data on specific hazards that are of interest to the region.

Varying considerations are necessary to plan for natural physical processes, man-made, and technological hazards. Even an individual hazard such as a flood can involve many physical processes that could be considered from views of hydrology, soil science, climate, meteorology, etc. Planning the resources and procedures needed to respond to specific aspects of hazards differs greatly; however, the structure for how operations are to occur can be consistent. A common structure for how operations are to occur can be planned by adopting an “all hazards approach” (FEMA 1996). It is important to have an all hazards policy to have a unified approach to handling disasters; having a structure of operations is necessary for decision-makers and responders to proceed in light of compounding events.

Often hazards have multiple types of impacts such as a hurricane which is a meteorological event often associated with the high winds involved, but resulting flooding can also be significantly impactful from both storm surge and rainfall. The March 2011 Japanese disaster was a compounding disaster case as it started with an earthquake which caused a tsunami and both of which contributed to a nuclear accident. Earthquakes are primarily caused by geologic shifts, and tsunamis are generated by geological and bathymetric configurations which can cause flooding. In this case in Japan, flooding compounded the event by making the backup diesel generators nonfunctional. Part of being resilient is having an adaptive system to face this compounding events as many failures are made worse by responses that are not flexible and previously good plans can cause unforeseen failures (David Woods 2010).

Consistent experience with a hazard can lead to general knowledge of what to plan for given an event. Tornadoes in Oklahoma or earthquakes in Japan are examples of hazards that are probable to

occur in specific areas so that the population gains familiarity. However, after many years without a major hazard of a specific type, the collective memory of knowing how to recognize such phenomena as a tsunami is lost. Unfortunately, many people went out on the beach when the waters receded during the 2004 Indian Ocean tsunami without realizing that this is an indicator to flee. Likewise, there is a risk of having too much experience as people can become complacent, for example, not seeking safety when they hear a tornado warning or not feeling the need to evacuate because the tsunami walls have handled smaller previous events and “the big one” is not foreseeable.

Conclusion

Disaster planning is essential to the success of further stages in the disaster management cycle. It is necessary to analyze where disasters are most likely to occur in order to be prepared for the event by having a data-driven understanding of human interests, physical attributes, and resources available. However, due to a perspective that there is less risk, there is often less data collection planning implemented for disasters that are unforeseen to occur in an area or to be so severe that it leads to unexpected challenges. Organized planning is needed at all levels of society for many different types of physical processes, man-made, and technological hazards which require unique planning considerations.

Cross-References

- [Big Geo-Data](#)
- [Big Variety Data](#)
- [Data Fusion](#)

Further Reading

FEMA. (1996). *Guide for all-hazard emergency operations planning*. Washington, DC: The Federal Emergency Management Agency. <https://www.fema.gov/pdf/plan/slg101.pdf>.

- Hultquist, C., Simpson, M., Cervone, G., and Huang, Q. (2015). Using nightlight remote sensing imagery and Twitter data to study power outages. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management (EM-GIS '15)*. ACM, New York, NY, Article 6, 6 pages. <https://doi.org/10.1145/2835596.2835601>.
- Woods, D. (2010). How do systems manage their adaptive capacity to successfully handle disruptions? A resilience engineering perspective. *Complex adaptive systems – Resilience, robustness, and evolvability: Papers from the Association for the Advancement of artificial intelligence (AAAI): Fall symposium (FS-10-03)*.

Discovery Analytics, Discovery Informatics

Connie L. McNeely

George Mason University, Fairfax, VA, USA

While big data is a defining feature of today's information and knowledge society, a huge and widening gap exists between the ability to accumulate the data and the ability to make effective use of it to advance discovery (Honavar 2014). This issue is particularly prominent in scientific and business arenas and, while the growth and collection of massive and complex data have been made possible by technological development, significant challenges remain in terms of its actual usefulness for productive and analytical purposes. Realizing the potential of big data to both accelerate and transform knowledge creation, discovery requires a deeper understanding of related processes that are central to its use (Honavar 2014). Advances in computing, storage, and communication technologies make it possible to organize, annotate, link, share, discuss, and analyze increasingly large and diverse data. Accordingly, aimed at understanding the role of information and intelligent systems in improving and innovating scientific and technological processes in ways that will accelerate discoveries, discovery analytics and discovery informatics are focused on identifying processes that require knowledge assimilation and reasoning.

Discovery analytics – involving the analysis and exploration of the data to determine trends and patterns – and discovery informatics – referring to the application and use of related findings – are based on the engagement of principles of intelligent computing and information systems to understand, automate, improve, and innovate various aspects of those processes (Gil and Hirsh 2012). Unstructured data is of particular note in this regard. Generated from various sources (e.g., the tracking of website clicks, capturing user sentiments from online sources or documents such as social media platforms, bulletin boards, telephone calls, blogs, or fora) and stored in nonrelational data repositories, or the “data lake,” vast amounts of unstructured data are analyzed to determine patterns that might provide knowledge insights and advantages in various arenas, such as business intelligence and scientific discovery. Discovery analytics are used to mine vast portions of the data lake for randomly occurring patterns; the bigger the data in the lake, the better the odds of finding random patterns that, depending on interpretation, could be useful for knowledge creation and application (Sommer 2019).

In reference to big data, discovery analytics has been delineated according to four types of discovery: visual, data, information, and event (Smith 2013; Cosentino 2013). Visual discovery has been linked, for example, to big data profiling and capacities for visualizing data. Combined with data mining and other techniques, visual discovery attends to enhanced predictive capability and usability. Data discovery points to the ability to combine and relate data from various sources, with the idea of expanding what is possible to know. Data-centric discovery is interactive and based on massive volumes of source data for analysis or modeling. Information discovery rests on search technologies, especially among widely distributed systems and big data. Different types and levels of search are core to information discovery based on a variety of sources from which big data are derived, from documents to social media to machine data. Event discovery – also called “big data in motion” – represents operational intelligence, involving the data collected on and observation of various phenomena,

actions, or events, providing rationales to explain the relationships among them.

In practical terms, given their critical roles in realizing the transformative potential of big data, discovery analytics and informatics can benefit multiple areas of societal priority and well-being (e.g., education, food, health, environment, energy, and security) (Honavar 2014). However, also in practical terms, discovering meaning and utility in big data requires advances in representations and models for describing and predicting underlying phenomena. Automation dictates the translation of those representations and models into forms that can be queried and processed. In this regard, computing – the science of information processing – offers tools for studying the processes that underlie discovery, concerned primarily with acquiring, organizing, verifying, validating, integrating, analyzing, and communicating information. Automating aspects of discovery and developing related tools are central to advancing discovery analytics and informatics and realizing the full potential of big data. Doing so means meeting challenges to understand and formalize the representations, processes, and organizational structures that are crucial to discovery; to design, develop, and assess related information artifacts; and to apply those artifacts and systems to facilitate discovery (Honavar 2014; Gil and Hirsh 2012; Dzeroski and Todorovski 2007).

Cross-References

- [Data Lake](#)
- [Data Mining](#)
- [Informatics](#)
- [Unstructured Data](#)

Further Reading

Cosentino, T. (2013, August 19). Three major trends in new discovery analytics. *Smart Data Collective*. <https://www.smartdatacollective.com/three-major-trends-new-discovery-analytics/#:~:text=One%20of%20those%20pillars%20is%20what%20is%20called,discovery%2C%20data%20discovery%2C%20informatics%20discovery%20and%20event%20discovery>.

Dzeroski, S., & Todorovski, L. (Eds.). (2007). *Computational discovery of communicable scientific knowledge*. Berlin: Springer.

Gil, Y., & Hirsh, Y. (2012). *Discovery informatics: AI opportunities in scientific discovery*. AAAI Fall Symposium Technical Report FS-12-03.

Honavar, V. G. (2014). The promise and potential of big data: A case for discovery informatics. *Review of Policy Research*, 31(4), 326–330.

Smith, M. (2013, May 7). Four types of discovery technology for using big data intelligently. <https://marksmith.ventanaresearch.com/marksmith-blog/2013/05/07/four-types-of-discovery-technology-for-using-big-data-intelligently>.

Sommer, R. (2019). Data management and the efficacy of big data: An overview. *International Journal of Business and Management*, 7(3), 82–86.

Diversity

Adele Weiner¹ and Kim Lorber²

¹Audrey Cohen School For Human Services and Education, Metropolitan College of New York, New York, NY, USA

²Social Work Convening Group, Ramapo College of New Jersey, Mahwah, NJ, USA

Diversity and Big Data

Diversity reflects a number of different sociocultural demographic variables including, but not limited to, race, ethnicity, religion, gender, national origin, disability, sexual orientation, age, education, and socioeconomic class. Big data refers to extremely large amounts of information that is collected and can be analyzed to identify trends, patterns, and relationships. The data itself is not as important as how it is used. Census data is an example of big data that provides information about characteristics across nations and populations. Other big data is used by multinational organizations, such as the World Bank and the United Nations, to help document and understand how policies and programs differentially affect diverse populations. In the USA, analysis of big data on voting, housing, and employments patterns led to the development of affirmative action and anti-discrimination policies

and laws that identify and redress discrimination based on diversity characteristics.

Self-reported Diversity Information

Many of the mechanisms used to create big datasets depend on self-reports as with the US Census and public school records. When self-reporting, individuals may present themselves inaccurately because of concerns about the use of the data or their perceived status within society. For example, self-descriptions of race or ethnicity may not be reported by respondents because of political or philosophical perceptions of inadequate categories, which do not meet an individual's self-definition. Some data such as *age* may appear to be fairly objective, but the person completing the form may have inaccurate information or other reasons for being imprecise. Options for identifying *sex* may only be *male* and *female*, which requires transgender and other individuals to select one or the other when, perhaps, they differently self-identify.

The data collection process or forms may introduce inaccuracies in the information. On the 2010 US Census short form, *race* and *ethnicity* seem to be merged. Korean, Chinese, and Japanese are listed as races. In the 2010 Census, questions ask about the relationship of each household member to person #1 (potentially the head of the household) and rather than *spouse* it has the choice of *husband* or *wife*. Many gay and lesbian individuals, even if married, may not use these terms to self-identify and hopefully *spouse* will be provided as an answer option in the next cycle. The identification of each person's sex on the form may currently allow the federal government to identify same-sex marriages. On the other hand, individuals who have concerns about privacy and potential discrimination may not disclose their marital status. Heterosexual couples, living as if they are married, may self-identify as such even if not legally wed. Census data is primarily descriptive and can be used by both municipalities and merchants to identify certain populations.

In 2010, the US Census eliminated the long-form, which was administered to only a sample of

the population and replaced it with the American Community Survey (ACS). The ACS is a continuous survey designed to provide reliable and timely demographic, housing, social, and economic data every year. This large dataset is much more extensive than that collected by the Census and offers the opportunity to determine the relationship between some diversity variables, such as gender, race, and ethnicity to economic, housing, employment, and educational variables. Again, this data is self-reported and persons completing the form may interpret questions differently. For example, one question asks the respondent about how well they speak English (very well, well, not well, not at all). It is easy to see how a native speaker of English and a person for whom it is a second language may have different fluency self-perceptions. In addition, it is possible to identify individuals with functional disabilities from this survey but not specifics.

Private and Public Records and Diversity

Both public and private information is aggregated into large datasets. Although individual health information is private, it is collected and analyzed by insurance networks and governmental entities. Health datasets may be used to make inferences about the health needs and utilization of services by diverse populations. Demographic data may demonstrate certain health conditions that are more prevalent among specific populations. For example, the Centers for Disease Control uses collected data on a variety of health indicators for African-Americans, Hispanic, or Latino populations, and men's and women's health. This data is grouped and provides information about the health needs of various populations, which can focus prevention, education, and treatment efforts.

The development of such large databases has been established by health networks and insurance companies to facilitate health care. In the USA, the Health Insurance Portability and Accountability Act (HIPPA) established rules regarding the use of this data and to protect the privacy of individuals' medical information. The

Center for Medicare and Medicaid has developed large datasets of information collected by health care providers that can be analyzed for research and policy and programming decisions.

Other records can be used to collect population information when paired with demographic diversity variables. School records can be used to highlight the needs of children in a given community. Library book borrowing is recorded and can provide information about the needs and interests of book borrowers. Data collected by the Internal Revenue Service for tax purposes can also be used to identify low-income neighborhoods. And certainly information collected by government programs such as Social Security, Medicare, Medicaid, Temporary Assistance for Needy Families (TANF), and the Supplemental Nutrition Assistance Program (Food Stamps) can link diversity to incomes, housing, and other community needs.

Social Media and Retail Data

Information about diversity can also be gleaned from a variety of indirect data collection methods used by social media and retail sources. The hair products a person buys may provide clues as to their race while the types of books they purchase may indicate religious beliefs. Retailers use this kind of information for targeted advertising campaigns and special offers for potential sales while also selling customer lists of relevant consumers to other vendors. This occurs in both online and brick and mortar stores when a person uses their credit cards. Medical equipment, cosmetics, foods and spices, books and vitamin supplements all may give clues to a person's race, age, religion, ethnicity, sexuality, or disability. When one uses a credit or store discount card the information is aggregated to create a profile of the consumer, even though they have not provided this information directly. Analysis of such large amounts of consumer information allows retailers to adapt their inventory to meet the specific needs of their customers and to individually market to them directly through electronic or mail promotions. For example, a person who buys cosmetics

primarily used by African-Americans might receive additional communications about a new line of ethnically diverse dolls.

Big data is generated when an individual searches the Internet, even if they do not purchase an item. Many free services, such as Facebook and Google, use analysis of page views to place targeted advertisements on members' pages. Not only do they earn income from these ads but when a person "Likes" an item, the advertisement is then showed to all the others in the person's network. Social media companies also use this data to show other products they estimate the user may be interested in. This can easily be demonstrated by this simple experiment. Go online and look at a variety of items not normally of interest to you and see how long it takes for advertisements of these products to appear on the webpages you visit. Imagine what happens if a person looks for sensitive information online, perhaps about sexuality, abortion, or a serious medical condition, and another family member uses the same device and sees advertisements linked to these private searches. This is even more challenging when people use work computers where there is no assurance of privacy.

Conclusion

People are becoming increasingly aware that current big data mining and analytics will provide private information to be used without their permission. There are concerns about the ways diversity identification information can be used by retailers, governments, and insurers. Such information can positively redress discrimination and inequities experienced by individuals who are members of diverse, minority groups. On the other hand, accessing this information may violate individual privacy. In the age of big data and the Internet, new data collection methods are being created and used in ways not covered by current legislation. Regulations to maintain privacy and prevent data from being used to discriminate against diverse groups need to be adjusted to deal with the rapidly changing access to data.

Cross-References

- [Biomedical Data](#)
- [Census Bureau \(U.S.\)](#)
- [Facebook](#)
- [Gender and Sexuality](#)
- [Google](#)
- [Religion](#)

Further Reading

American Community Survey – <http://www.census.gov/acs/www/>.
 Centers for Disease Control and Prevention – Health Data Interactive – <http://www.cdc.gov/nchs/hdi.htm>.
 Population Reference Bureau – <http://www.prb.org/>.
 The World Bank – Data – <http://data.worldbank.org/>.
 The United Nations – UNdata – <http://data.un.org/>.
 U.S. Census – <http://www.census.gov/> – <http://data.un.org/>.

Document-Oriented Database

- [NoSQL \(Not Structured Query Language\)](#)

DP

- [Data Processing](#)

Driver Behavior Analytics

Seref Sagioglu
 Department of Computer Engineering, Gazi
 University, Ankara, Turkey

Driver or Driving Behavior Analytics

Internet technologies support many new fields of research, innovation, technology as well as developing new applications and implementations. Recently, big data analytics and technologies might help to improve quality, system,

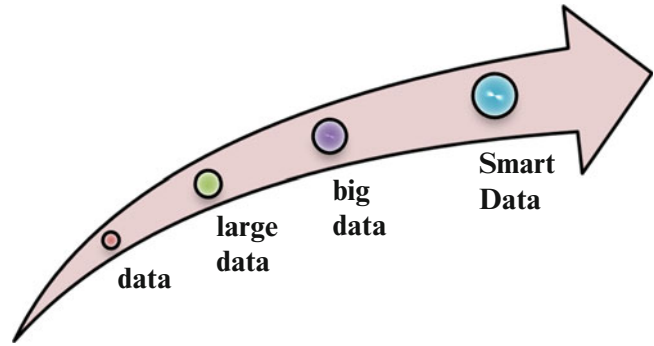
production, process, progress, and productivity in many fields, institutions, sectors, applications or implementations. They also help to make better plan and decision, to give better service, to take advantage, to establish new company, to get new discovery, to have new output, finding, perception as well as thought, even judgment with the support of big data techniques, technologies and analytics. In order to analyse, model, establish, forecast or predict driver/driving behaviour, driver in duty (man, woman, machine), driving media (in vehicle system) and driving environment (inside or outside vehicle) are considered. In order to understand these explanations clearly, first thing to do is to understand the data, data types, data volume, data structure, method and methodology in data analysis and analytics. Figures 1 and 2 briefly show the revolution and change of data journey. If these data are collected properly, better analysis or analytics can be achieved. More benefits, outcomes, findings and profits might be acquired by industry, sector, university or institution.

In order to understand big data analytics, it is important to understand the concept of driving/driver behavior to develop better and faster systems. As shown in Fig. 3, to explain and analyze driver/driving behavior, there are three major issues considered as given below:

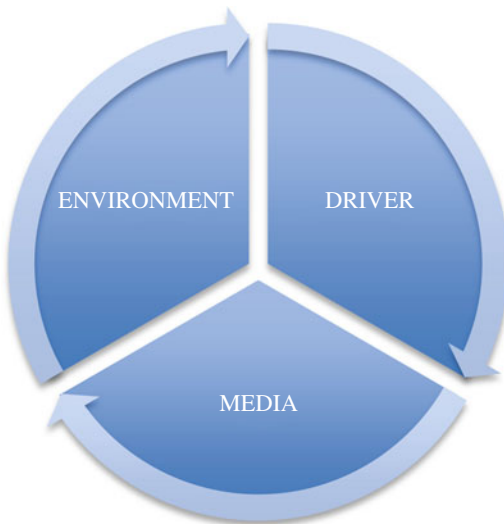
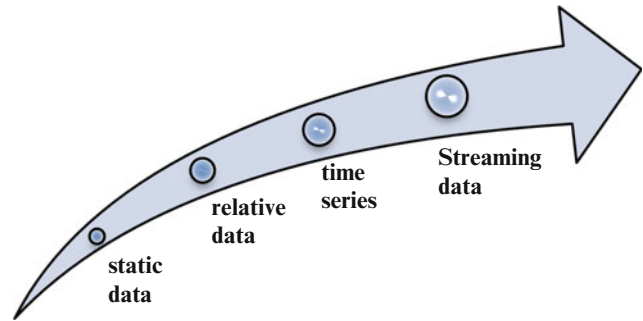
1. Driver (man, woman, machine, etc.)
2. Driving media (car, lorry, track, motobike, cycle, ship, flight, etc.)
3. Driving environment (roads, motorways, highways, inner city, intersections, heavy traffics, highway design standards, crowds, weather conditions, etc.)

Driving behavior or driver behavior is a cyclic process covering media, environment, and driver as illustrated in Fig. 3. Any value can be achieved analyzing the data acquired from drivers, driving media, and driving environments especially using big data analytics. Even if driving behavioral data were categorized into three different groups in the literature (Miyajima et al. 2006; Wakita et al. 2005), it can be categorized into five considering big data analytics for better understanding. Big

Driver Behavior Analytics, Fig. 1 Data revolution



Driver Behavior Analytics, Fig. 2 Data types



Driver Behavior Analytics, Fig. 3 Basic dynamics of driving behavior

1. Vehicle operational data:
 - Steering angle, velocity, acceleration, engine speed, engine on-off, etc.
2. Vehicle data:
 - Gas pedal position, Various sensor data, maintenance records, etc.
3. Driver psychiatric/psychologic data:
 - Driver record, driver mood, chronic illness, hospital records, etc.
4. Driver physical data:
 - Usage or driving records, following distance, drowsiness, face parts, eye movement, etc.
5. Vehicle outside data:
 - Outside temperature, humidity, distance for other vehicles, speeds of other vehicles, road signs, radio alerts, heavy traffic record, coordinate info, etc.

data features play a crucial role due to availability of different types of vehicles, sensors, environments, and technology support. The categorization covering new suggestions is given below as:

As can be seen above, there have been many data types available in the literature. It can be clearly pointed out that most of the studies in the literature have focused on the vehicle data and vehicle operational data. It is expected that all of five data types, or even more, might be used for

modeling driver behavior considering big data analytics.

Driver/driving behavior can be estimated or predicted with the help of available models, formulas, or theories. When the literature is reviewed, there have been many methods based on machine learning, SVM, Random Forrest, Naive Bayes, KNN, K-means, statistical methods, MLP, Fuzzy neural network (FNN), Gaussian mixture model, and HMM models applied to modeling, predicting, and estimating driver behaviors (Enev et al. 2016; Kwak et al. 2017; Wakita et al. 2006; Meng et al. 2006; Miyajima et al. 2007; Nishiwaki et al. 2007; Choi et al. 2007; Wahab et al. 2009; Dongarkar and Das 2012; Van Ly et al. 2013; Zhang et al. 2014). It should be emphasized that there are no big data analytics solution yet.

In order to understand the mathematics behind this topic, some of the articles and important models available in the literature are reviewed and summarized.

– **A model for car following task** (Wakita et al. 2005):

This task is a car following (stimulus response) model and involves following a vehicle with a constant distance in front, and adjusting the relative velocity as stimuli is to calculate the response of drivers by either accelerating or decelerating action as in (Wakita et al. 2005).

$$\dot{v}(t+T) = C_1 \dot{h}(t) + C_2 \{h(t) - D\}$$

where

C_1 or C_2 is the response sensitivity to the stimulus.

D is the optimum distance to the vehicle in front.

T is the response delay. These values may be the constants or the functions of other variables.

– **Cepstral Analysis for Gas Pedal (GP) and Brake Pedal (BP) Pressure** (Öztürk and Erzin 2012; Campo et al. 2014):

Cepstral feature extraction is used for driving behaviour signals as reported in (Öztürk

and Erzin 2012; Campo et al. 2014). The selected signals, GP pressure and BP pressure are evaluated for each frame k (the short term real cepstrum) and K cepstral features f are extracted in (Öztürk and Erzin 2012) as:

$$f_k = F^{-1} \text{BPF} \{ \log F(x(n + kT)) \}$$

where

$x(n + kT)$: the frame signal multiplied by the window,

BPF: the band-pass filter separating noise from driving behavior signals,

F : denotes the discrete-time Fourier transform (DTFT), and

F^{-1} : denotes its inverse.

– **Spectral analysis of pedal signals** (Öztürk and Erzin 2012):

It can be seen that the spectra are similar in the same driver but it is different among two drivers. Assuming that the spectral envelope can capture the differences in between the characteristics of different drivers (Öztürk and Erzin 2012).

– **GMM driver modeling and identification** (Jensen et al. 2011).

A Gaussian mixture model (GMM) (Jensen et al. 2011) was used to represent the distributions of feature vectors of cepstrum of each driver. The GMM parameters were estimated using the expectation maximization (EM) algorithm. The GMM driver models were evaluated in driver identification experiments, in which the unknown driver was identified as driver k who gave the maximum weighted GMM log likelihood over gas pedal and brake pedals:

$$k = \text{argmax} \{ A \log P(\text{GP} | \lambda_{G,k}) + (1 - A) \log P(\text{BP} | \lambda_{B,k}) \}$$

where

$$0 \leq A \leq 1.$$

GP and BP are the cepstral sequences of gas and brake pedals.
 $\lambda_{G,k}$ and $\lambda_{B,k}$ are the k-th driver models of GP and BP, respectively.
A is the linear combination weight for the likelihood of gas pedal signals.

In order to have data for driving analysis, simulators or data generators were used for data collection (Wakita et al. 2006; Meng et al. 2006; Miyajima et al. 2007; Zhang et al. 2014), but today most data have been picked up from real environments mounted on/in vehicles, carried mobile devices, or wearable devices on drivers. Especially, driving behavioral signals are collected using data collection vehicles designed and supported by companies, projects, and research groups (Miyajima et al. 2006; Hallac et al. 2016; Enev et al. 2016; Kwak et al. 2017; Nishiwaki et al. 2007; Choi et al. 2007; Wahab et al. 2009; Dongarkar and Das 2012; Van Ly et al. 2013; Zhang et al. 2014). It should be emphasized that recent researches focus on data collection from vehicles via CAN or other protocols for further complex analysis in many other fields.

When the literature is reviewed, there have been a number of approaches to analyze driver behaviors. The features or parameters of data collected from vehicles are given in Table 1. These data are obtained from the literature (Miyajima et al. 2006; Hallac et al. 2016; Wakita et al. 2005; Enev et al. 2016; Kwak et al. 2017; Hartley 2000; Colten and Altevogt 2006; Jensen et al.

2011; Salemi 2015; Öztürk and Erzin 2012; Campo et al. 2014; Wakita et al. 2006; Meng et al. 2006; Miyajima et al. 2007; Nishiwaki et al. 2007; Choi et al. 2007; Wahab et al. 2009; Dongarkar and Das 2012; Van Ly et al. 2013; Zhang et al. 2014) and combined for representation.

Today technology supports to acquire those parameters and data given above from many internal and external sensors. These sensors might be multi-sensors, audio, video, picture or text. In some cases, questionnaires are also used for this analysis. The literature on driver behavior also covers available techniques and technologies (Miyajima et al. 2006; Hallac et al. 2016; Wakita et al. 2005; Enev et al. 2016; Kwak et al. 2017; Hartley 2000; Colten and Altevogt 2006; Jensen et al. 2011; Salemi 2015; Öztürk and Erzin 2012; Campo et al. 2014; Wakita et al. 2006; Meng et al. 2006; Miyajima et al. 2007; Nishiwaki et al. 2007; Choi et al. 2007; Wahab et al. 2009; Dongarkar and Das 2012; Terzi et al. 2018; Van Ly et al. 2013; Zhang et al. 2014). When the analysis based on perception of big data is considered, more parameters and data types might be used to analyze driver behavior more accurately and compactly. For doing that, the data not only collected in Table 1 but also the data such as weather condition, road safety info, previous health or driving records of drivers, accident records, road condition, real-time alerts for traffic, traffic jam, speed, etc. can be also used for big data analytics for achieving better models, evaluations, results,

Driver Behavior Analytics, Table 1 Parameters used for estimating/predicting/modeling driver behavior

- Vehicle speed, acceleration, and deceleration - Steering, steering wheel - Gear shift - Engine condition, torque, rpm, speed, coolant temperature - Vehicle air conditioning - Yaw rate - Shaft angular velocity - Fuel consumption - Mass air flow rate	- Brake pedal position, pressure - Gas (accelerator) pedal position, pressure - Transmission oil temperature, activation of air compressor, torque converter speed, wheel velocity, rear, front, left hand, right hand - Retrader - Throttle position - Start-stop - Turning signal	- Face mood - Head movement - Sleepy face, sleepiness, tiredness, - Gyro - Stress, drowsiness - Lane deviation - Long-term fuel trim bank, intake air pressure, friction torque, calculated load value - Following distance from vehicle ahead
--	---	--

outcomes or values. Especially, a recent and comprehensive survey introduced by Terzi et. al. (2018) provides a recent big data perspective on driver / driving behavior and discusses the contribution of big data analytics to the automotive industry and reserach field.

Conclusions

This entry concludes that driver behavior analytics is a challenging problem. Even if there have been many studies available in the literature, the studies having big data analytics are very rare. In order to achieve this task, the points are discussed and given below.

Developing a study on driver behavior based on big data analytics:

- Requires suitable infrastructure, algorithms, platforms, and enough data for better and faster analytics
- Enables more or better models for driver behavior
- Provides solution not only on one driver but also a large number of drivers belonging to a company, institution, etc.
- Requires not only data but also smart data for further analytics
- Provides new solutions, gains, or perceptions for problems
- Needs experts and expertise for getting expected solutions
- Costs more than classical approaches

There are other issues that might affect the success or failure of big data analytics for modeling/predicting driving behavior due to:

- Availability of not enough publications and big data sources for researches.
- Collecting proper data having different data sets, time intervals, size, format, or parameters.
- Limitations of bandwidth of mobile technologies or operators used in transferring data from vehicles to the storage to collect or to analyze the data for further progresses.

- Benefit-cost relations. It should be considered that in some cases the cost would be high in comparison with the achieved value. It should be emphasized that having big data means does not always guarantee to get values from this analytics.
- Facing lost connection in some places to transfer the data from vehicle to the system.
- The lack of algorithms needed to be used in real-time applications.

Final words, the solutions and suggestions provided in this chapter might help to reduce traffic accidents, injuriousness, loss, traffic jams, etc. and also increase productivity, quality, safety not only for safer, better and comfortable driving but also designing, developing and manufacturing better vehicles, establishing better roads and comfortable driving, etc.

Further Reading

- Campo, I., Finker, R., Martinez, M. V., Echanobe, J., & Doctor, F. (2014). *A real-time driver identification system based on artificial neural networks and cepstral analysis*, 2014 I.E. International Joint Conference on Neural Networks (IJCNN), 6–11 July 2014, Beijing, pp. 1848–1855.
- Choi, S., Kim, J., Kwak, D., Angkititrakul, P., & Hansen, J. H. (2007). Analysis and classification of driver behavior using in-vehicle can-bus information. In *Biennial workshop on DSP for In-vehicle and mobile systems*, pp. 17–19.
- Colten, H. R., & Altevogt, B. M. (Eds.). (2006). *Sleep disorders and sleep deprivation, an unmet public health problem. : Institute of medicine (US) committee on sleep medicine and research*. Washington, DC: Natioenal Academies Press (US). ISBN: 0-309-10111-5.
- Dongarkar, G. K., & Das, M. (2012). Driver classification for optimization of energy usage in a vehicle. *Procedia Computer Science*, 8, 388–393.
- Enev, M., Takakuwa, A., Koscher, K., & Kohno, T. (2016). Automobile driver fingerprinting. In *Proceedings on Privacy Enhancing Technologies*, 2016(1), 34–50.
- Hallac, D., Sharang, A., Stahlmann, R., Lamprecht, A., Huber, M., Roehder, M., Sosic, R., & Leskovec, J. (2016). *Driver identification using automobile sensor data from a single turn 2016 I.E. 19th international conference on intelligent transportation systems (ITSC) Windsor Oceanico Hotel, Rio de Janeiro, 1–4 Nov 2016*.

- Hartley, L. (2000). *Review of fatigue detection and prediction technologies*. Melbourne: National Road Transport Commission.
- Jensen, M., Wagner, J., & Alexander, K. (2011). Analysis of in-vehicle driver behaviour data for improved safety. *International Journal of Vehicle Safety*, 5(3), 197–212.
- Kwak, B. I., Woo, J. Y., & Kim, H. K. (2017). Know your master: Driver profiling-based anti-theft method, arXiv:1704.05223v1 [cs.CR] 18 April.
- Meng, X., Lee, K. K., & Xu, Y. (2006). Human driving behavior recognition based on hidden markov models. In *IEEE International Conference on Robotics and Biomimetics 2006 (ROBIO'06)* (pp. 274–279). Kunming: China.
- Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., & Takeda, K. (2006). Cepstral analysis of driving behavioral signals for driver identification. In *IEEE International Conference on ICASSP*, 14–19 May 2006, (pp. 921–924). Toulouse: France. <https://doi.org/10.1109/ICASSP.2006.1661427>.
- Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., & Itakura, F. (2007). Driver modeling based on driving behavior and its evaluation in driver identification. In *Proceedings of the IEEE*, 95(2), 427–437.
- Nishiwaki, Y., Ozawa, K., Wakita, T., Miyajima, C., Itou, K., & Takeda, K. (2007). Driver identification based on spectral analysis of driving behavioral signals. In J. H. L. Hansen and K. Takeda (Eds.), *Advances for in-vehicle and mobile systems - 2007 Challenges for International Standards* (pp. 25–34). Boston: Springer.
- Öztürk, E., & Erzin, E. (2012). Driver status identification from driving behavior signals. In J. H. L. Hansen, P. Boyraz, K. Takeda, and H. Abut (Eds.), *Digital Signal Processing for In-Vehicle Systems and Safety*. Springer NY, pp. 31–55.
- Salemi, M. (2015). Authenticating drivers based on driving behavior, Ph.D. dissertation. Rutgers University, Graduate School, New Brunswick.
- Terzi, R., Sagioglu, S., & Demirezen, M. U. (2018) Big data perspective for driver/driving behavior. In *IEEE Intelligent Transportation Systems Magazine*, Accepted for Publication.
- Van Ly, M., Martin, S., & Trivedi, M. M. (2013). Driver classification and driving style recognition using inertial sensors. In *IEEE Intelligent Vehicles Symposium (IV)*, 23–26 June 2013, (pp. 1040–1045). Gold Coast: Australia.
- Wahab, A., Quek, C., Tan, C. K., & Takeda, K. (2009). Driving profile modeling and recognition based on soft computing approach. In *IEEE Transactions on Neural Networks*, 20(4), 563–582.
- Wakita, T., Ozawa, K., Miyajima, C., Igarashi, K., Itou, K., Takeda, K., & Itakura, F. (2005). Driver Identification Using Driving Behavior Signals. In *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*, Vienna, 13–16 Sept 2005.
- Wakita, T., Ozawa, K., Miyajima, C., Garashi, K. I., Katunobu, I., Takeda, K., & Itakura, F. (2006). Driver identification using driving behavior signals. *IEICE Transactions on Information and Systems*, 89(3), 1188–1194.
- Zhang, X., Zhao, X., & Rong, J. (2014). A study of individual characteristics of driving behavior based on hidden markov model. *Sensors & Transducers*, 167(3), 194.

Drones

R. Bruce Anderson^{1,2} and Alexander Sessums²

¹Earth & Environment, Boston University,
Boston, MA, USA

²Florida Southern College, Lakeland, FL, USA

In the presence of the “information explosion” Big Data means big possibilities. Nowhere is the ultimacy of this statement more absolute than in the rising sector of Big Data collecting via unmanned aerial vehicles or drones. The definition of Big Data, much like its purpose, is all encompassing, overarching, and umbrella-like in function. More fundamentally, the term refers to the category of data sets which are so large they prove standard methods of data interrogation to no longer be helpful or applicable. In other words, it is “a large volume unstructured data which cannot be handled by standard data base management systems like DBMS, RDBMS, or ORDBMS.” Instead, new management software tools must be employed to “capture, curate, manage and process the data within a tolerable elapsed time.” The trend toward Big Data in the past few decades has been in large measure due to analytical tools which allows experts to now spot correlations in large data volumes where they would have previously been indistinguishable and therefore more accurately identify trends in solutions.

One technique used to grow data sets is through the technique of remote sensing and the use of aerial drones. A drone is an unmanned aerial vehicle which can be operated in two distinct ways: either it can be controlled autonomously by onboard software or wirelessly by grounded personnel. What was once a closely guarded military venture has now been

popularized by civilian specialists and is now producing a sci-fi like sector of the Big Data economy. Over the past two decades, a dramatic rise in the number of civilian drone applications has empowered a breathtaking number of civilian professionals to utilize drones for specialist tasks such as nonmilitary security, firefighting, photography, agricultural and wildlife conservation. While the stereotypical drone is usually employed on tasks that are too “dull, dirty or dangerous” for human workers to attempt, drones extend themselves to seemingly endless applications. What has been coined, “the rise of the drones” has produced an amount of data so vast it may never be evaluated and indeed, critics claim it was never intended to be evaluated but stored and garnished as a prized information set with future monetary benefit. For example, on top of the mountains of photographs, statistics, and surveillance the United States military’s Unmanned Aircraft Vehicles (UAVs) have amassed over the past few years, drone data is phenomenally huge and growing ever larger. The age of big drone data is here and its presence is literally pressing itself into our lives, leaving a glut of information for analysts to decipher and some questioning its ethical character. What was once a fairy-tale concept imagined by eccentric data managers, is now an accessible and comprehensible mass of data.

Perhaps the greatest beneficiary of the drone data revolution is in the discipline of agriculture. Experts predict that 80% of the drone market will be dedicated to agriculture over the next 10 years, resulting in what could potentially be a \$100 billion dollar industry by the year 2025. The technique of drones feeding data sets on the farm is referred to as Precision Agriculture and is quickly becoming a favorite tool of farmers. Precision Agriculture is now allowing farmers greater access to statistical information as well as the ability to more accurately interpret natural variables throughout the life cycle of their crop ecosystems. In turn, Precision Agriculture enables farmers increased control over day-to-day farm management and increases farmer’s agility in reacting to market circumstances. For example, instead of hiring an aviation company to perform daily pest application, a farmer can

now enlist the capabilities of a preprogrammed drone to apply precise amounts of pesticide in precise locations without wasting chemicals. In the American mid-western states, where the economic value of farming is tremendous, tractor farming has now become intelligent. GPS laden planting tools now enable farmers to “monitor in real-time-while they’re planting-where every seed is placed.” Drones have immense potential for the farming community and provide intelligent capabilities for smart data collection. Perhaps the chief capability being the ability to identify farm deficiencies and then compute data sets into statistical solutions for farmers. Precision data sets function “by using weather reports, soil conditions, GPS mapping, water resources, commodity market conditions and market demand” and allows “predictive analytics” to then be applied to “improve crop yield” and therefore encourage farmers to “make statistical-based decisions regarding planting, fertilizing and harvesting crops.” This anywhere, anytime accessibility of drones is incredibly attractive to modern farmers and is offering tremendous improvement to under-producing farms. As small, cost-effective drones begin to flood American farms, drones and data sets will offer a boon of advantages that traditional farming methods simply cannot compete against.

Drones and Big Data are also offering themselves to other civilian uses such as meteorology and forestry. Many are expecting that drones and “Big data will soon be used to tame big winds.” That is because the American science agency, National Oceanic and Atmospheric Administration, recently released a pack of “Coyote” drones for use in capturing data on hurricanes. The three-foot drone will enable data to be collected above the surface of the water and below the storm, a place previously inaccessible to aviators. This new access will enable forecasters to better analyze direction, intensity, and pressure of incoming hurricanes well before they reach land, allowing for better preparations to be formed. Drones have also found employment in forestry fighting wild-fires. Traditional methods of firefighting were often based on “paper maps and gut feelings.” With drones, uncertainty is reduced allowing for

“more information for less cost and it doesn’t put anyone in harm’s way.”

In the military world, especially within Western superpowers, Big Data is now supporting military ventures and enabling better combat decisions to be made during battle. However, big military data is now posing big problems that the civilian world has yet to encounter – the military has a “too much data problem.” Over the past decade thousands of terabytes of information have been captured by orbiting surveillance drones from thousands of locations all across the world. Given the fact that the use of drones in conventional warfare coincided with the height of American activity in the Middle East, the American military complex is drowning in information. Recently, the White House announced it would be investing “more than \$200 million” in six separate agencies to develop systems that could “extract knowledge and insights from large and complex collections of digital data.” It is thought that this investment will have big advantages for present and future military operations.

Working in tandem, drones and Big Data offer tremendous advantages for both the civilian and military world. As drones continue to become invaluable information gatherers, the challenge will be to make sense of the information they collect. As data sets continue to grow, organization and application will be key. Analysts and interpretation software will have to become increasingly creative in order to decipher the good data from the menial. However, there is no doubt that drones and Big Data will play a big part in the future of ordinary lives.

Further Reading

- Ackerman, S. (2013, April 25). Welcome to the age of big drone data. Retrieved September 1, 2014.
- Bell, M. (n.d.). US Drone Strikes are Controversial are they war crimes. Retrieved September 1, 2014.
- Big Data: A Driving Force in Precision Agriculture. (2014, January 1). Retrieved September 1, 2014.
- CBS News – Breaking News, U.S., World, Business ... (n.d.). Retrieved September 1, 2014.
- Lobosco, K. (2013, August 19). Drones can change the fight against wildfires. Retrieved September 1, 2014.

Noyes, K. (2014, May 30). Cropping up on every farm: Big Data technology. Retrieved September 1, 2014.

Press, G. (2015, May 9). A very short history of Big Data. Retrieved September 1, 2014.

Wirthman, L. (2014, July 28). How Drones and Big Data are creating winds of change for Hurricane Forecasts. Retrieved September 1, 2014.

Drug Enforcement Administration (DEA)

Damien Van Puyvelde
University of Glasgow, Glasgow, UK

Introduction

The Drug Enforcement Administration (DEA) is the lead US government agency in drug law enforcement. Its employees investigate, identify, disrupt, and dismantle major drug trafficking organizations (DTOs) and their accomplices, interdict illegal drugs before they reach their users, arrest criminals, and fight the diversion of licit drugs in the United States and abroad. Formally a part of the Department of Justice, the DEA is one of the largest federal law enforcement agencies with close to 10,000 employees working domestically and abroad. In recent years, the DEA has embraced the movement towards greater use of big data to support its missions.

Origins and Evolution

The US government efforts in the area of drug control date back to the early twentieth century when the Internal Revenue Service (IRS) actively sought to restrict the sale of opium following the passage of the Harrison Narcotics Act of 1914. The emergence of a drug culture in the United States and the expansion of the international drug market led to further institutionalization of drug law enforcement in the second half of the twentieth century. The US global war on the manufacture, distribution, and use of narcotics started when Congress passed the Controlled Substances

Act of 1970, and President Richard Nixon established the DEA. Presidential reorganization plan no. 2 of 1973 merged pre-existing agencies – the Bureau of Narcotics and Dangerous Drugs (BNDD), the Office for Drug Abuse Law Enforcement (ODALE), and the Office of National Narcotics Intelligence (ONNI) – into a single agency. This consolidation of drug law enforcement sought to provide momentum in the “war on drugs,” better coordinate the government’s drug enforcement strategy, and make drug enforcement more accountable.

Less than a decade after its 1982 inception, Attorney General William French Smith decided to reorganize drug law enforcement in an effort to centralize drug control and to increase the resources available for the “war on drugs.” Smith gave concurrent jurisdiction to the Federal Bureau of Investigation (FBI) and the DEA, and while the DEA remained the principal drug enforcement agency, its administrator was required to report to the FBI director instead of the associate attorney general. This arrangement brought together two of the most important law enforcement agencies in the United States and inevitably generated tensions between them. Such tensions have historically complicated the implementation of the US government’s drug control policies.

Over the past 40 years, the DEA has evolved from a small, domestic-oriented agency to one primarily concerned with global law enforcement. In its early days, the DEA employed some 1470 special agents for an annual budget of \$74 million. Since then, DEA resources have grown steadily, and, in 2014, the agency employed 4700 special agents, 600 diversion investigators, over 800 intelligence research specialists, and nearly 300 chemists, for a budget of 2.7 billion USD.

Missions

Although the use of drugs has varied over the last four decades, the issues faced by the DEA and the missions of drug law enforcement have remained consistent. Put simply, DEA’s key mission is to

put drug traffickers in jail and to dismantle their conspiracy networks. The DEA enforces drug laws targeting both illegal drugs, such as cocaine and heroin, and legally produced but diverted drugs including stimulants and barbiturates. One of its major responsibilities is to investigate and prepare for the prosecution of major violators of controlled substance laws that are involved in the growing, manufacturing, and distribution of controlled substances appearing in or destined for illicit traffic in the United States. When doing so, the DEA primarily targets the highest echelons of drug trafficking by means of the so-called kingpin strategy. It is also responsible for the seizure and forfeiture of assets related to illicit drug trafficking. From 2005 to 2013, the DEA stripped drug trafficking organizations of more than \$25 billion in revenues through seizures. A central aspect of the DEA’s work is the management of a national drug intelligence program in cooperation with relevant agencies at the federal, state, local, and international levels. The DEA also supports enforcement-related programs aimed at reducing the availability and demand of illicit controlled substance. This includes the provision of specialized training for state and local law enforcement. It is also responsible for all programs associated with drug law enforcement counterparts in foreign countries and liaison with relevant international organizations on matters relating to international drug control.

The agency has four areas of strategic focus related to its key responsibilities. First, international enforcement includes all interactions with foreign counterparts and host nations to target the leadership, production, transportation, communications, finance, and distribution of major international drug trafficking organizations. Second, domestic enforcement focuses on disrupting and dismantling priority target organizations, that is to say, the most significant domestic and international drug trafficking and money laundering organizations threatening the United States. Third, the DEA advises, assists, and trains state and local law enforcement and local community groups. Finally, the agency prevents, detects, and eliminates the diversion of controlled substances that are diverted to the black market and whose

use may be diverted from their intended (medical) use.

With the advent of the Global War on Terrorism in 2001, the DEA has increasingly sought to prevent, disrupt, and defeat terrorist organizations. In this context, narcoterrorism – which allows hostile organizations to finance their activities through drug trafficking – has been a key concern. This nexus between terrorism and drug trafficking has effectively brought the DEA closer to the US Intelligence Community.

Organization

The DEA headquarters are located in Alexandria, VA, and the agency has 221 domestic offices organized in 21 divisions throughout the United States. The DEA currently has 86 offices in 67 countries around the world. Among these foreign offices, a distinction is made between the more important country office and smaller resident and regional offices. The DEA is currently headed by an administrator and a deputy administrator who are both appointed by the president and confirmed by the senate. At lower levels, DEA divisions are led by special agents in charge or SAC.

The agency is divided into six main divisions: human resources, intelligence, operations, operational support, inspection, and financial management. The DEA is heavily dependent on intelligence to support its missions. Its intelligence division collects information from a variety of human and technical sources. Since the DEA is primarily a law enforcement agency, intelligence is primarily aimed at supporting its operations at the strategic, operational, and tactical levels. DEA analysts work in conjunction with and support special agents working in the field; they search through police files and financial and other records and strive to demonstrate connections and uncover networks. Throughout the years, DEA intelligence capabilities have shifted in and out of the US Intelligence Community. Since 2006, its Office of National Security Intelligence

(ONSI) is formally part of the Intelligence Community. ONSI facilitates intelligence coordination and information sharing with other members of the US Intelligence Community. This rapprochement can be seen as an outgrowth of the link between drug trafficking and terrorism. The operations division conducts the field missions organized by the DEA. Special agents plan and implement drug interdiction missions and undercover operations and develop networks of criminal informants (CI). The cases put together by analysts and special agents are often made against low-level criminals and bargained away in return for information about their suppliers, in a bottom-up process. Within the operations division, the special operations division coordinates multi-jurisdictional investigations against major drug trafficking organizations. The majority of DEA special agents are assigned to this branch, which forms a significant part of the agency as a whole. A vast majority of DEA employees are special agents, while analysts form a minority of less than a thousand employees. The prominence of special agents in the agency reflects its emphasis on action and enforcement. The operational support division provides some of the key resources necessary for the success of the DEA's mission, including its information infrastructure and laboratory services. Within DEA laboratories, scientific experts analyze seized drugs and look for signatures, purity, and information on their manufacturing and the routes they may have followed.

The government's effort to reduce the supply and demand of drugs is a broad effort that involves a host of agencies. Many of the crimes and priority organizations targeted by the DEA transcend standard drug trafficking, and this requires effective interagency coordination. Within the federal government, the DEA and the FBI have primary responsibility for interior enforcement, which concerns those organizations and individuals who distribute and use drugs within the United States. The shared drug law enforcement responsibilities between these two agencies were originally supposed to fuse DEA

street knowledge and FBI money laundering investigation skills, leading to the establishment of joint task forces. Other agencies, such as the Internal Revenue Service (IRS), Immigration and Customs Enforcement (ICE), and Bureau of Customs and Border Patrol (CBP), are also involved in drug law enforcement. For example, IRS assists the DEA with the financial aspects of drug investigations, and CBP intercepts illegal drugs and traffickers at entry points to the United States. Since the DEA has sole authority over drug investigations conducted abroad, it cooperates with numerous foreign law enforcement agencies, providing them with assistance and training to further US drug policies. Coordination between government agencies at the national and international levels continues to be one of the main challenges faced by DEA employees as they seek to implement US drug laws and policies.

Criticisms

The DEA has been criticized for focusing excessively on the number of arrests and seizures it conducts each year. The agency denied approximately \$25.7 billion in drug trafficking revenues through the seizure of drugs and assets from 2005 to 2013, and its arrests rose from 19,884 in 1986 to 31,027 in 2015. Judging by these numbers, the agency has fared well in the last decades. However, drug trafficking and consumption have risen consistently in the United States, and no matter how much the agency is arresting and seizing, the market always provides more drugs. Targeting networks and traffickers has not deterred the widespread use of drugs. Some commentators argue that this focus is counterproductive to the extent that it leads to further crimes, raises the cost of illicit drugs, and augments profit of drug traffickers. Other criticisms have focused on the ways in which the DEA targets suspects and have accused the agency of engaging in racial profiling.

Critiques hold that drug control should focus more on the demand than the supply side and on the reasons behind the consumption of illegal

drugs rather than on drug traffickers. From this perspective, the agency's resources would be better spent on drug treatment and education. Although the DEA has made some efforts to tackle demand, the latter have remained limited. The DEA's "kingpin strategy" and its focus on hard drugs like heroin and cocaine have also been criticized for their ineffectiveness because they overlook large parts of the drug trafficking business.

From Databases to Big Data

The ability to access, intercept, collect, and process data is essential to combating crime and protecting public safety. To fight the "war on drugs," the DEA developed its intelligence capabilities early on, making use of a centralized computer database to disseminate and share intelligence. The DEA keeps computer files on millions of persons of interest in a series of databases such as the Narcotics and Dangerous Drug Information System (NADDIS). In the last few decades, DEA investigations have become increasingly complex and now frequently require sophisticated investigative techniques including electronic surveillance and more extensive document and media exploitation, in order to glean information related to a variety of law enforcement investigations. The increasing volumes and complexity of communications and related technologies have been particularly challenging and have forced the law enforcement agency to explore new ways to manage, process, store, and disseminate big databases. When doing so, the DEA has been able to rely on private sector capabilities and partner agencies such as the National Security Agency (NSA). Recent revelations (Gallagher 2014) suggest that the DEA has been able to access some 850 billion metadata or records about phone calls, emails, cellphone locations, and Internet chats, thanks to a search engine developed by the NSA. Such tools allow security agencies to identify investigatory overlaps, track suspects' movements, map out their social

networks, and make predictions in order to develop new leads and support ongoing cases.

The existence of large databases used by domestic law enforcement agencies poses important questions about the right to privacy, government surveillance, and the possible misuse of data. Journalists reported that telecommunication companies' employees have worked alongside DEA agents to supply them with phone data based on records of decades of American phone calls. These large databases are reportedly maintained by telecommunication providers, and the DEA uses administrative subpoenas, which do not require the involvement of the judicial branch, to access them. This has fostered concerns that the DEA may be infringing upon the privacy of US citizens. On the whole, the collection and processing of big data is only one aspect of the DEA's activities.

Cross-References

- [Data Mining](#)
- [National Security Agency \(NSA\)](#)
- [Social Network Analysis](#)

Further Reading

- Gallagher, R. (2014). The surveillance engine: How the NSA built its own secret Google. *The Intercept*. <https://firstlook.org/theintercept/2014/08/25/icreach-nsa-cia-secret-google-crisscross-proton/>. Accessed 5 Apr 2016.
- Lyman, M. (2006). *Practical drug enforcement*. Boca Raton: CRC Press.
- Van Puyvelde, D. (2015 online). Fusing drug enforcement: The El Paso intelligence center. *Intelligence and National Security*, 1–15.
- U.S. Drug Enforcement Administration. (2009). *Drug enforcement administration: A tradition of excellence, 1973–2008*. San Bernardino: University of Michigan Library.

E

E-agriculture

► AgInformatics

Earth Science

Christopher Round

George Mason University, Fairfax, VA, USA

Booz Allen Hamilton, Inc., McLean, VA, USA

Earth science (also known as geoscience) is the field of sciences dedicated to understanding the planet Earth and the processes that impact it, including the geologic, hydrologic, and atmospheric sciences (Albritton and Windley 2019). Geologic science concerns the features and composition of the solid Earth, hydrologic science refers to the study of Earth's water, and atmospheric science is the study of the Earth's atmosphere. Earth science aims to describe the planet's processes and features to understand its present state and how it may have appeared in the past, and will appear in the future.

Much of related research relies on Earth analytics, referring to the branch of data science used for Earth science. Big data in earth science is generated by satellites, models, networks of

sensors (which are often part of the Internet of Things), and other sources (Baumann et al. 2016; Yang et al. 2019), and data science is critical for developing models of complex Earth phenomena. Time-series and spatial elements are common attributes for Earth science data. In reference to big data, technologies such as cloud computing and artificial intelligence are now being used to address challenges in using earth science data for projects that historically were difficult to conduct (Yang et al. 2019). While big data is used in traditional statistical analysis and model development, increasingly machine learning and deep learning are being utilized with big data in Earth analytics for understanding nonlinear relationships (Yang et al. 2019).

Earth science contributes to and interacts with other scientific fields such as environmental science and astronomy. As an interdisciplinary field, environmental science incorporates the earth sciences to study the environment and provide solutions to environmental problems. Earth science also contributes to astronomy, focused on celestial objects, by contributing information that could be valuable for the study of other planets. In general, research and knowledge from earth science contribute to other sciences and play significant roles in understanding and acting on global issues (including social and political issues). For example, the race to access resources in the Arctic is a result of the understanding of changes in the

cryosphere (the portion of the earth that is solid ice), which is receding in response to rises in global temperatures (IPCC 2014; Moran et al. 2020).

In this regard, much of our understanding of climate change and future projections come from complex computer models reliant on big data synthesized from a wide variety of data inputs (Farmer and Cook 2013; Schnase et al. 2016; Stock et al. 2011). These models' granularity has been tied to available computer processing power (Castro 2005; Dowlatabadi 1995; Farmer and Cook 2013). Considerable time must be placed into justifying the assumptions in these models and into what they mean for decision makers in the international community. On a more local level, weather reports, earthquake predictions, mineral exploration, etc. are all supported by the use of big data and computer modeling (Dastagir 2015; Hewage et al. 2020; Kagan 1997; Sun et al. 2019).

Further Reading

- Albritton, C. C., & Windley, B. F. (2019, November 26). *Earth sciences*. Encyclopedia Britannica. <https://www.britannica.com/science/Earth-sciences>
- Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., Bigagli, L., Boldrini, E., Bruno, R., Calanducci, A., Campalani, P., Clements, O., Dumitru, A., Grant, M., Herzig, P., Kakaletis, G., Laxton, J., Koltida, P., Lipskoch, K., et al. (2016). Big data analytics for earth sciences: The EarthServer approach. *International Journal of Digital Earth*, 9(1), 3–29. <https://doi.org/10.1080/17538947.2014.1003106>.
- Castro, C. L. (2005). Dynamical downscaling: Assessment of value retained and added using the regional atmospheric modeling system (RAMS). *Journal of Geophysical Research*, 110(D5). <https://doi.org/10.1029/2004JD004721>.
- Dastagir, M. R. (2015). Modeling recent climate change induced extreme events in Bangladesh: A review. *Weather and Climate Extremes*, 7, 49–60. <https://doi.org/10.1016/j.wace.2014.10.003>.
- Dowlatabadi, H. (1995). Integrated assessment models of climate change: An incomplete overview. *Energy Policy*, 23(4–5), 289–296.
- Farmer, G. T., & Cook, J. (2013). Types of models. In G. T. Farmer & J. Cook (Eds.), *Climate change science: A modern synthesis: volume 1—The physical climate* (pp. 355–371). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-5757-8_18.
- Hewage, P., Trovati, M., Pereira, E., & Behera, A. (2020). Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*. <https://doi.org/10.1007/s10044-020-00898-1>.
- IPCC. (2014). IPCC, 2014: Summary for policymakers. In *Climate change 2014: Impacts, adaptation, and vulnerability. Part a: Global and sectoral aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.
- Kagan, Y. Y. (1997). Are earthquakes predictable? *Geophysical Journal International*, 131(3), 505–525. <https://doi.org/10.1111/j.1365-246X.1997.tb06595.x>.
- Moran, B., Samso, J., & Feliciano, I. (2020, December 12). *Warming arctic with less ice heats up Cold War tensions*. PBS NewsHour. <https://www.pbs.org/newshour/show/warming-arctic-with-less-ice-heats-up-cold-war-tensions>
- Schnase, J. L., Lee, T. J., Mattmann, C. A., Lynnes, C. S., Cinquini, L., Ramirez, P. M., Hart, A. F., Williams, D. N., Waliser, D., Rinsland, P., Webster, W. P., Duffy, D. Q., McInerney, M. A., Tamkin, G. S., Potter, G. L., & Carriere, L. (2016). Big data challenges in climate science: Improving the next-generation cyber-infrastructure. *IEEE Geoscience and Remote Sensing Magazine*, 4(3), 10–22. <https://doi.org/10.1109/MGRS.2015.2514192>.
- Stock, C. A., Alexander, M. A., Bond, N. A., Brander, K. M., Cheung, W. W. L., Curchitser, E. N., Delworth, T. L., Dunne, J. P., Griffies, S. M., Haltuch, M. A., Hare, J. A., Hollowed, A. B., Lehoudey, P., Levin, S. A., Link, J. S., Rose, K. A., Rykaczewski, R. R., Sarmiento, J. L., Stouffer, R. J., et al. (2011). On the use of IPCC-class models to assess the impact of climate on living marine resources. *Progress in Oceanography*, 88(1), 1–27. <https://doi.org/10.1016/j.pocean.2010.09.001>.
- Sun, T., Chen, F., Zhong, L., Liu, W., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. *Ore Geology Reviews*, 109, 26–49. <https://doi.org/10.1016/j.oregeorev.2019.04.003>.
- Yang, C., Yu, M., Li, Y., Hu, F., Jiang, Y., Liu, Q., Sha, D., Xu, M., & Gu, J. (2019). Big earth data analytics: A survey. *Big Earth Data*, 3(2), 83–107. <https://doi.org/10.1080/20964471.2019.1611175>.

Eco-development

► Sustainability

E-Commerce

Lázaro M. Bacallao-Pino
University of Zaragoza, Zaragoza, Spain
National Autonomous University of Mexico,
Mexico City, Mexico

Synonyms

[Electronic commerce](#); [Online commerce](#)

Electronic commerce, commonly known as e-commerce, has been defined in several ways, but, in general, it is the process of trading – both buying and selling – products or services using computer networks, such as the Internet. Although a timeline for the development of e-commerce usually includes some experiences during the 1970s and the 1980s, analyses agree in considering that it has been from the 1990s onward when there has been a significant shift in methods of doing business with the emergence of the of e-commerce. It draws on a diverse repertory of technologies, from automated data collection systems, online transaction processing, or electronic data interchange to Internet marketing, electronic funds transfer, and mobile commerce, usually using the WWW for at least one process of the transaction's life cycle, although it may also use other technologies such as e-mail.

As a new way of conducting businesses online, e-commerce has become a topic analyzed by academics and businesses, given its rapid growth – some estimates consider that global e-commerce will reach almost \$1.4 trillion in 2015, while Internet retail sales for 2000 were \$25.8 billion – and the increasing trend in “consumers” purchasing decisions to be made in an online environment, as well as the raising number of people engaging in e-commerce activities. The world's largest e-commerce firms are the Chinese Alibaba Group Holding Ltd., with sales for 2014 estimated at \$420 billion; Amazon, with reported sales of

\$74.4 billion for 2013; and eBay, with reported sales of \$16 billion for 2013. Other major online providers with a strong presence in their home and adjacent regional markets are Rakuten in Japan, Kobo in India, Wuaki in Spain, or Zalando in Europe.

Among the far-reaching ramifications of the emergence of the Internet as a tool for the business-to-consumer (B2C) aspect of e-commerce, an aspect underlined by many analyses is the necessity to understand how and why people participate in e-commerce activities, in a context where businesses have more opportunities to reach out to consumers in a very direct way. Regarding this aspect, the novelty of the online environment and, consequently, the e-commerce as a phenomenon has produced a diversity of criteria for understanding online shopping behavior, from positions that consider actual purchases as a measure of shopping in this case to others that employ self-reports of time online and frequency of use as a criterion.

When analyzing consumer behavior in e-commerce activities, two dimensions highlighted by many studies are customer loyalty and website design. On the one hand, changes in customer loyalty in e-commerce have been a topic of particular concern among researchers and businessmen, since the instantaneous availability of information in the Internet has been seen as a circumstance that vanishes brand loyalty as potential buyers can compare the offerings of sellers worldwide, reducing the information asymmetries among them and modifying the bases of customer loyalty in digital scenarios. On the other hand, website design is considered one of the most important factors for successful user experiences in e-commerce. Besides the website's usability – recognized as a key to e-commerce success – several researchers have also argued that a successful e-commerce website should be designed in a way that inspires “customers” trust and engagement, persuading them to buy products. It is assumed that its elements affect online consumers' intentions to the extent of even influencing their beliefs related to e-commerce and, consequently, their

attitudes as e-buyers, as well as their sense of confidence or perceived behavioral controls.

In that scenario of an excess of information, e-commerce websites have developed some technological resources to facilitate consumers' online decisions by giving them information about product quality and some assistance on product search and selection. One of those technological tools is the so-called recommendation agents (RAs), a software that – based on individual consumers' interests and preferences about products – provides certain advice on products that match these interests and predilections. RAs then become technological resources that can potentially improve the quality of consumers' decisions by helping to reduce the amount of information they have about products as well as the complexity of the online search in e-commerce websites.

As specifically e-commerce technological resources, RAs set a number of debates associated with some critical social issues, including privacy and trust. Those applications are part of the efforts of commercial websites to provide certain information about a product to attract potential online shoppers, but as in this case, those suggestions are based on some information collected from users – their preferences, shopping history or browsing pattern, or the pattern of choices by other consumers with similar profiles – then consumers are concerned about what information is collected, whether it is stored, and how it is used, mainly because it can be obtained explicitly but also in an implicit way. Seen by some authors as an example of mass customization in e-commerce, RAs include different models or forms, from general recommendation lists or specific and personalized suggestions of products to customer comments and ratings and community opinions or critiques to notification services and other deep kinds of personalizations.

Debates on a Multidimensional Phenomenon

Debates on e-commerce have highlighted both its opportunities and challenges for businesses. For many authors, advantages of e-commerce include

aspects such as the millions of products available for consumers at the largest e-commerce sites, the access to narrow market segments that are widely distributed, greater flexibility, lower cost structures, faster transactions, broader product lines, greater convenience, as well as better customization. The improvement of the quality of products and the creation of new methods of selling have also been considered benefits of e-commerce. But, at the same time, it has been noted that e-commerce also sets a number of challenges for both consumers and businesses, such as choosing among the many available options, the consequences of the new online environment for the processes of buying and selling – for instance, the particularities of “consumers” online decision making – consumers' confidence, and privacy or security issues.

E-commerce is a market strategy where enterprises may or may not have a physical presence. Precisely, the online/offline interrelationships have been a particular point at issue in the analyses of e-commerce. For instance, some authors have analyzed its impact on shopping malls that includes a change in shopping space, rental contracts, the shopping mall visit experience, service, image, multichannel strategy, or lower in-shop prices. Instead of regarding e-commerce as a threat, these analyses suggest that shopping malls should examine and put in practice integration strategies with the e-commerce, for instance, by virtual shopping malls, portals, click and collect, check and reserve, showrooms, or virtual shopping walls.

In the same line, another dimension of analyses has been the comparisons between returns obtained by conventional firms from e-commerce initiatives and returns to the net firms that only have online presence. Other relevant issues have been the analysis of how do returns from business-to-business (B2B) e-commerce compare with returns from B2C e-commerce and how do the returns to e-commerce initiatives involving digital goods compare to initiatives involving tangible goods. In this sense, there are opposite opinions; while some authors have considered that the opportunities in the B2B e-commerce field far exceed the opportunities in B2C one, others have suggested that the increasing role of ICTs in everyday life

becomes raising opportunities for the B2C e-commerce.

Tendencies on E-Commerce

Some general assumptions on e-commerce have suggested that, as a consequence of the reduced search costs associated with the Internet, it would encourage consumers to abandon traditional marketplaces in order to find lower prices and online sellers would be more efficient than offline competitors, forcing traditional offline stores out of business. Other authors have considered that, since there are increasing possibilities of direct relationships between manufacturers and consumers, some industries would become disintermediated. However, contrary to those tendencies, previous researches have noted that few of those assumptions proved to be correct since the structure of retail marketplace in countries with high levels of B2C e-commerce – such as the United States – has not followed those trends, because consumers also give importance to other aspects besides prices, such as brand name, trust, reliability, and delivery time.

The main trends in e-commerce, observed by different studies, include a tendency towards social shopping users, who share their opinions and recommendations with other buyers through online viral networks, in line with the increasing use of interactive multimedia marketing – blogs, user-generated content, video, etc. – and the Web 2.0. At the same time, analyses agree in noting the increasing profits of e-commerce, as well as the raising diversity of goods and services available online and the average annual amount of purchases. The development of customized goods and services, the emphasis on improved online shopping experience by focusing on easy navigation or offering on-line inventory updates, and the effective integration of multiple channels by sellers –including alternatives such as on-line order and in-store pickup, “bricks-and-clicks,” “click and drive,” online web catalog, gift cards, or in-store web kiosk ordering – are also some relevant trends.

E-commerce innovation and changes are, to some extent, inherently associated to the

permanently changing nature of ICTs and the continuous development of new technologies and applications, following a process through which sellers’ strategies and consumer behavior evolve with the technology. In that sense, two current tendencies are the emergence of what has been called social commerce and increasing mobile e-commerce or m-commerce.

Social commerce, a new trend with no stable and agreed-upon definition that has been a topic of research for few analyses, refers to the evolution of e-commerce as a consequence of the adoption of Web 2.0 resources to enhance customer participation and achieve greater economic value. Debates on it analyze, for instance, specific design matters of social commerce platforms – such as Amazon and Starbucks on Facebook – and their relations to e-commerce and Web 2.0 and propose a multidimensional model of it that includes individual, conversation, community, and commerce levels.

Mobile e-commerce, for its part, although was originally proposed in the late 1990s for referring to the delivery of e-commerce capabilities into the consumer’s hand via wireless technologies, has recently become a subject of debate and research, as the number of smartphones raised and this is the primary way for one-third of smartphone users for going online. M-commerce has moved away from SMS systems and into current applications, avoiding this way security vulnerabilities and congestion problems. Many payment methods are available for m-commerce consumers, including premium-rate phone numbers, charges added to the consumer’s mobile phone bill and credit cards –allowing, in some cases, credit cards to be linked to a phone’s SIM card – micropayment services, or stored-value cards, frequently used with mobile device application stores or also music stores. Some authors argue that this transition to m-commerce will have similar effects as the Internet had for traditional retailing in the late 1990s.

Although early m-commerce consumers appear to be previous heavy e-commerce users, researches on the social dimension of the phenomenon have concluded, e.g., that there are differences in user behavior across the mobile applications and regular Internet sites and, besides

this, mobile shopping applications appear to be also associated to an immediate and sustained increase in total purchasing. These findings on m-commerce corroborate the articulation between the technological, business, and sociocultural and behavioral dimensions of e-commerce.

Cross-References

- [Information Society](#)
- [Online Advertising](#)

Further Reading

- Einav, L., Levin, J., Popov, I., & Sundaresan, N. (2014). Adoption, and use of mobile E-commerce. *American Economic Review: Papers & Proceedings*, 104(5), 489–494. <https://doi.org/10.1257/aer.104.5.489>.
- Huang, Z., & Benyoucef, M. (2013). From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications*, 12(4), 246–259.
- Laudon, K.C., & Guercio Traver, C. (2008). *E-commerce: Business, technology, society*. Upper Saddle River: Pearson Prentice Hall.
- Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5(1), 115–153.
- Turban, E., Lee, J. K., King, D., Peng Liang, T., & Turban, D. (2009). *Electronic commerce 2010*. Upper Saddle River: Prentice Hall Press.

Economics

Magdalena Bielenia-Grajewska and
Magdalena Bielenia-Grajewska
Division of Maritime Economy, Department of
Maritime Transport and Seaborne Trade,
University of Gdansk, Gdansk, Poland
Intercultural Communication and
Neurolinguistics Laboratory, Department of
Translation Studies, University of Gdansk,
Gdansk, Poland

Economics can be briefly defined as the discipline that focuses on the relation between

resources, demand and supply of individuals and organizations, as well as the processes that are connected with the life cycle of products. Walter Wessels (2000) in his definition highlights that economics shows people how to allocate their scarce resources. For centuries, people have been making economic choices about the most advantageous process of allocating relatively scarce resources and choosing the needs to be met. From this perspective, economics is the science of how people use the resources at their disposal to meet various material and non-material needs. However, big data have brought dramatic changes to economics as a field. In particular, beyond traditional econometric methods, new analytic skills and approaches – especially those associated with machine learning – are required to engage big data for economics research and applications (Harding and Hersh 2018).

Processes of Rational Management (Change from Homo Oeconomicus to the Machine of Economics)

According to D.N. Wagner (2020), changing economics practice includes the process of discovering how economic patterns change under the influence of technological innovations. He claims that one of the specific economic pattern influenced by artificial intelligence (AI) is the so-called *machina economica* (its predecessor was *homo oeconomicus*) entering the world economy. What is more, Wagner (2020) shows that disciplines like economics and computer science use an analytical perspective rooted in institutional economics. In more details, the author presents the economic model in the world with AI using an analytical angle grounded in institutional economics, in the context of artificial intelligence, it is no surprise that artificial intelligence agents were also created as economic actors. Moreover, he claims that homo economicus has long served as a desirable role model for artificial intelligence (AI). The first researcher interested in the economic rationality of man was A. Smith who introduced the model of man rationally operating in the

sphere of economy. According to the ideology of man as an individual seeking to maximize profit, it is worth noting that an entrepreneur can be treated as *homo oeconomicus*. The paradigm of mainstream economics postulates that the economic entity (*homo oeconomicus*) is guided by its own interest when making decisions. The perspective of perceiving economic man's decisions is called methodological individualism. This position assumes that individual egoism (selfish motives; internally defined interest) is of great importance because the decisions of countless free individuals create social welfare. Moreover, mainstream economics proves that the economic system is the sum of all the economic units (*homo oeconomicus*). Epistemology based on methodological individualism ignores the very important fact that an individual making free choices acts in a specific social context, it does not remain in isolation from the surrounding world. At the beginning of the institutional changes in the years 1980–1990, the paradigm of economism was used as an effect of looking for an answer to the question about the proper paradigm of socio-economic development. The characteristics of the economics paradigm in economic sciences are discussed based on the *homo oeconomicus* model, with economic decisions based on the economic value of results. The mainstream economists perceive many successes, including economic development in the economic unit. Taking into account the recent economic crisis, it is worth remembering that the mainstream economic paradigm should be enriched with contextual analysis (social, cultural, etc.). The new institutional economics (NIE) is criticized for reducing human existence to *homo oeconomicus*, excluding its social rooting. The criticism about the achievements of mainstream economics is caused by some questionable assumptions of using the concept of *homo oeconomicus*, seen as a model of rational choice in its extreme version, and by the assumption of the rules of the game as a result of the interaction of individuals. As D.N. Wagner claims, institutional economic perspective and influence of neoclassical economics (with model of man – *homo oeconomicus* – the so-called welfare role model for AI) establish suitable notions

and analytical frameworks for a world with artificial intelligence.

The observation of economic reality shows that complementing the economic analysis typical of mainstream economics with the theme of methodological holism becomes almost essential. Institutional economists (the advocates of the new institutional economy) see this issue in a similar way as sociologists who reject the assumption of perfect rationality of individual subjects (individuals) making decisions. A broader and more multi-faceted concept is needed. The new institutional economy examines socio-economic phenomena much better than neoclassical economics, mainly by assuming limited individual rationality. To be predictive, the set of fundamental assumptions (paradigm) of modern economics should be based on the assumption of the emergence of levels of integration of social phenomena (the assumption of a new institutional economy). In this sense, the new institutional economy is making a statement of two perspectives: methodological individualism and methodological holism. Methodological individualism has been described as the position of mainstream economics that preaches the primacy of *homo oeconomicus*, while economic theory, called the new institutional economy, refers to the need for methodological individualism and methodological holism, in which the starting point is the sociological man (*homo sociologicus*) or the socio-economic man. The theoretical orientation of methodological holism will be briefly and generally presented, within the framework of which the host entity should be perceived as an interconnected environment, thus showing that its decisions are influenced by historical, cultural and social context (the primacy of a holistic approach to phenomena). Epistemology based on methodological holism takes as its starting point the behavior of society (of a certain community and not the behavior of an individual) for understanding socio-economic mechanisms. The methodology called holism is undoubtedly functional in the process of analyzing the network of dependencies of a given community, because it allows showing the individual with the so-called blessing of the inventory and taking into account

the network of human relations with the institutional biospheres.

Economics: Different Typologies and Subtypes

As Bielenia-Grajewska (2015a) discusses, economics focuses on the problem of how scarce resources with different uses are distributed (allocated) in a society. The purpose of these activities is to produce goods. Since resources are limited, the possibilities of their use are numerous and diverse, therefore, it is examined how the produced goods are divided among members of society. Taking into account different typologies of economics discussed by researchers, one of the most well-known classifications of economics is the taxonomy of microeconomics and macroeconomics. Macroeconomic analysis allows to analyze the problems of allocation at the level of the whole society, the whole national economy. As far as macroeconomics is concerned, such issues as inflation, unemployment, demand and supply, business cycles, exchange rates as well as fiscal and monetary policy are examined by researchers. Microeconomic analysis helps consider allocation processes at the level of individual economic entities (enterprise, consumer). Microeconomics focuses on, among others, price elasticity, competition, monopoly and game theory.

Another division of economics takes into account the scope of research: national economics and international economics. The difference in focus is also visible in subcategorizing economics, by taking into account a sphere of life it governs; consequently, sports economics or economics of leisure or tourism can be distinguished. Economics is also studied through the type of economy, with such subtypes analyzed as market economy and planned economy. Although economics may seem for some as a traditional and fixed domain as far as the scope of research is concerned, it actively responds to the changes taking place in modern research. For example, the growing interest in neuroscience and cognition has led to the creation and development of such disciplines as behavioral economics and

neuroeconomics, studying how the brain and the nervous system may provide data on economic decisions. It should be stated that economics does not exist in a vacuum; economic conditions are determined by culture, politics, natural resources, etc. In addition, economics is not only a domain that is shaped by other disciplines as well as different internal and external factors, but it is also the type of discipline that influences other areas of life. For example, economics affects linguistics since it leads to the creation and dissemination of new terms denoting the economic reality.

Goods: Products and Services

In the process of management man produces goods to satisfy his needs. In economic theory, goods are divided into products and services. Products are those goods which by purchasing and using them people become their legal owners (e.g., food, clothing, furniture, car). Services, on the other hand, are economic goods giving purchasers the right to use them only temporarily (e.g., the service of airplane flight does not make a human being the owner of an airplane). In the era of the developing world economy there is a huge number of various products and services aimed at satisfying more and more exorbitant human needs. Together with technological developments, new products and services are constantly appearing, the manufacturers of which aim at satisfying more and more diverse and sophisticated needs of buyers.

In the area of big data, diverse data related to buyers and products are created, stored and analyzed. It is assumed that the concept of Big Data was first discussed in 1997 by Cox and Ellsworth in *Managing Big Data for Scientific Visualization*. The above-mentioned authors noted that there are two different meanings of this concept, such as big data collections and big data objects. The use of Big Data's and Data Science's knowledge and techniques is becoming increasingly common from the perspective of these products and services as well. Big Data, including the analysis of large data sets generated by various types of IT

systems, is widely used in various areas of business and science, including economics. The data streams generated by smartphones, computers, game consoles, appliances, household appliances, installations/software in apartments and homes, and even clothes – are of great importance for modern business. Emails, the location of mobile devices, social media, such as Facebook, LinkedIn, Twitter, or blogs, lead to the growth of data. Every activity taken on the Internet generates new information. From the perspective of goods and services, Data Science enables, on the basis of data sent from the above-mentioned devices, to determine users' preferences and even to forecast their future behavior. Big Data is a potential flywheel for the development of global IT, which is very important for the economy. Another issue concerning the use of the possibilities of Big Data is related to the demand for goods and services. The demand depends on the purchasing power (derived from the economic opportunity) of citizens. The concept of purchasing power is closely linked to the income of people and the price of goods. The demand for goods is treated as the amount of goods that people want to have if there is no limit to purchasing power. Technological development has enabled people to communicate more effectively; Big Data systems processing unimaginable amounts of data provide information for more efficient distribution of resources and products. Big Data (a large amount of data and its processing) over the next few years can become a tool to precisely target the needs of customers. Nowadays, the so-called wearable technology or wearable devices, i.e., intelligent electronic equipment that individuals wear, are very popular among users. The history of wearable technology began with a watch tracking human activity. Wearable devices are an example of the Internet of Things; the huge potential of the Internet of Things is also demonstrated by the current and forecast number of devices connected to the network. Due to the very dynamic development of the concept and application of Big Data in various areas of human activity, there is more and more talk about the possibility of using related methods of data analysis in activities aimed at improving competitiveness. Big Data is

considered a tool that will allow targeting of customer needs and forecasting investments, company portfolios, etc., with high accuracy.

Economy and Technological Development

As Bielenia (2020) states, the global expansion of online technologies has changed global business and the work environment for organization. The impact of Big Data in the economic field (data analysis and predictive modeling) is tremendous. Big Data encompasses a series of concepts and activities related to the acquisition, maintenance, and operation of data. It is worth mentioning that findings of EMC Digital Universe with Research & Analysis by IDC The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things proved that digital bits are doubling in size every 2 years. In 2020 it reached the value of 44 zettabytes, or 44 trillion gigabytes. Analyzing and utilizing Big Data leads to improved predictions. New databases and statistical techniques open up many opportunities.

Big Data has become an important part of economists' work and require new skills in computer science and databases. Thanks to the use of Big Data, analysts gain access to huge amounts of reference and comparative data that allow them to simulate social and economic processes. The increasing time range of the available data allows generating more and more reliable information about trends in the economy. Big Data is a tool that helps entities to better understand their own environment and the consumers who use their products or services. Big Data applied to the economy corresponds to the use of scarce resources. Doug Laney of META Group (now Gartner) in the publication 3D Data Management: Controlling Data Volume, Velocity, and Variety, defined the concept of Big Data in the "3 V" model: volume (the amount of data), velocity (the speed at which data is generated and processed), and variety (the type and nature of data). Over the years, the 3 V model has been expanded with an additional dimension veracity, creating a 4 V model. Extracting volume from

data is characterized by “4 Vs”: occurrence in large amount (volume), huge variety, high variability (velocity), and a significant value. According to Balar and Chaabita, Balar and Naji, Hilbert Big Data is characterized by 5Vs – volume, variety, velocity, value, and veracity. Volume refers to the sheer volume of generated and stored data. Variety means that the data comes from various sources. Big Data can use structured as well as unstructured data. Velocity corresponds to the speed at which the data arrives and the time it is analyzed. Value is related to data selection analyzed, i.e., which data will be relevant and valuable, and which will be useless. Finally, veracity relates to data reliability. In other words, data credibility relates to the truthfulness of the data as well as the data regularity.

Economics and Big Data

Taylor, Schroeder, and Meyer (2014) state that providing the definition of big data in economics is not an easy task. First of all, they stress that the discussion on the big versus not big data is still going on in social science. Secondly, economics as a discipline has been using databases and searching for tools that enable to deal with a considerably large amount of data for years. As Bielenia and Podolska (2020) state, an innovative economy based on knowledge and modern technological solutions could not function without the Internet. Technological development has meant that the generation of more and more computer data. There is a lot of data provided over the Internet and they come from various sources (market data, social networks, own sales systems or partner systems). The amount of data collected is enormous and grows with each new action performed via Internet by users. The concept of Big Data is therefore characterized by a few dimensions like volume, velocity, variety, value, and veracity. Most of the studies in the field of computer science that deal with the 3 V, 4 V, or 5 V problem in the context of managing and drawing knowledge from big data sets have one goal – how to tame Big Data and give it a structure. Thus, instead of providing a clear-cut definition,

the focus should be rather placed on some tendencies that determine the link between economics and big data.

Although economists have been facing studies of extensive amounts of data for years, modern economics has to deal with big data connected with rapid technological advancements in the sphere of economic activities. For example, nowadays many individuals purchase goods and services online and, consequently, e-commerce is connected with generating and serving big data related to sales, purchases, customers, and the role of promotion in customer decisions. Big data is also visible in the sphere of online trading, with many investors purchasing and selling financial instruments in the web. The birth of such financial opportunities as forex, continuous trading, or computerized stock exchanges have also influenced the amount of big data handled on an everyday basis. In addition, the relation between economics with big data can be observed from the discipline perspective. Thus, the role of big data may be studied by taking into account economics subdisciplines and their connection with big data. In microeconomics, big data is related with labor economics, whereas macroeconomists focus on big data related to monetary policies. Big data has also led to the growing research interest in econometrics and statistics that may facilitate the process of gathering and analyzing data. Taking into account the relatively new subdomains of economics, such as behavioral economics and neuroeconomics, neuroscientific tools facilitate the process of acquiring information (Bielenia-Grajewska 2013, 2015b). These tools prove to be useful especially if obtaining data from other sources is connected with the high risk of, e.g., respondents providing fake answers or leaving the questions unanswered. In addition, neuroscientific tools offer data on different issues simultaneously by observing the brain in a complex way. Apart from the scientific perspective, big data is also connected with studying economics. The learning dimension encompasses how data is managed by students and how the presented data influences the perception and cognition of economic notions. The profit dimension of economics being connected with big data is the focus on

the economical character of gathering and storing data. Thus, specialists should rely on the methods that do not generate excessive costs. Moreover, some companies try to sell big data they possess in order to gain profit.

Taking into account available methods, the performance of companies in relation to big data is studied by applying, e.g., the *Big Data Business Model Maturity Index* by Bill Schmarzo. Schmarzo's model consists of the following phases: Business Monitoring, Business Insights, Business Optimization, Data Monetization, and Business Metamorphosis. *Business Monitoring* encompasses the use of Business Intelligence and traditional methods to observe business performance and this stage concentrates on trends, comparisons, benchmarks and indices. The second phase, *Business Insights*, is connected with using statistical methods and data mining tools to deal with unstructured data. In the third phase called *Business Optimization*, the focus is on the automatic optimization of business operations. An example includes using algorithms in trading by financial companies. The fourth stage named *Data Monetization* is devoted to making an advantage of big data to generate revenue. An example is the creation of "intelligent products" that follow customer behaviors and needs. The last phase called *Business Metamorphosis* is connected with changing the company into a business entity operating in new markets or offering new services able to meet complex needs of customers. The size and changing parameters (unstructured) make traditional management and analysis impossible. As Racka (2016) claims, technological solutions used for very large datasets reaching huge sizes, vertical scaling (purchase of better and better machines for Big Data purposes), or horizontal scaling (expansion by adding more machines) are used. The advantage of Big Data technology is that it allows one to analyze the fast incoming and changing data in real time, without having to enter it into databases. Referring to the author, the most commonly used Big Data technology solutions currently include NoSQL, MapReduce, and Apache Hadoop. Also, Blazquez and Domenech (2018) define the data lifecycle within a Big Data paradigm. The steps to

manage data are as follows: (1) Study and planning, (2) Data collection, (3) Data documentation and quality assurance, (4) Data integration, (5) Data preparation, (6) Data analysis, (7) Publishing and sharing, and (8) Data storage and maintenance.

Cross-References

- Behavioral Analytics
- Business
- Decision Theory
- E-commerce

Further Reading

- Balar, K., & Chaabita, R. (2019). Big Data in economic analysis: Advantages and challenges. *International Journal of Social Science and Economic Research*, 04 (07).
- Balar, K., & Naji, A. (2015). A model for predicting ischemic stroke using data mining algorithms IJISSET. *International Journal of Innovative Science, Engineering & Technology*, 2(11).
- Bielenia, M. (2020). Different approaches of leadership in multicultural teams through the perspective of actor-network theory. In I. Williams (Ed.), *Contemporary applications of actor network theory*. Singapore: Palgrave Macmillan.
- Bielenia, M., & Podolska, A. (2020). Powszechny dostęp do Internetu jako prawo człowieka i warunek rozwoju gospodarczego. In B. Daria, K. Ryszard, & M. Prawnicze (Eds.), *Prawa człowieka i zrównoważony rozwój: konwergencja czy dywergencja idei i polityki*. Warszawa: Wydawnictwo C.H. Beck.
- Bielenia-Grajewska, M. (2013). International neuromanagement. In D. Tsang, H. H. Kazeroony, & G. Ellis (Eds.), *The Routledge companion to international management education*. Abingdon: Routledge.
- Bielenia-Grajewska, M. (2015a). Economic growth and technology. In M. Odekon (Ed.), *The SAGE encyclopedia of world poverty*. Thousand Oaks: SAGE Publications.
- Bielenia-Grajewska, M. (2015b). Neuroscience and learning. In R. Gunstone (Ed.), *Encyclopedia of science education*. Dordrecht: Springer.
- Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130.
- Cox, M., & Ellsworth, D. (1997). Managing Big Data for scientific visualization. Siggraph. www.dcs.ed.ac.uk/teaching/cs4/www/visualisation/SIGGRAPH/giga_byte_datasets2.pdf.

- EMC. <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>
- Harding, M., & Hersh, J. (2018). Big Data in economics. *IZA World of Labor*, 451. <https://doi.org/10.15185/izawol.451>.
- Hilbert, M. Big Data for development: A review of promises and challenges. *Development Policy Review*. martinhilbert.net. Retrieved 2015-10-07.
- Laney Doug. (2001). 3D data management: Controlling data volume, velocity, and variety. META Group (now Gartner) [online <http://blogs.gartner.com>].
- Racka, K. (2016). *Big Data – znaczenie, zastosowania i rozwiązania technologiczne*. Zeszyty Naukowe PWSZ w Płocku Nauki Ekonomiczne, t. XXIII.
- Schmarzo, B. (2013). *Big Data: Understanding how data powers big business*. Indianapolis: John Wiley & Sons, Inc.
- Taylor, L, Schroeder, R., & Meyer, E. (2014, July–December). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society*.
- Wagner, D. N. (2020). Economic patterns in a world with artificial intelligence. *Evolutionary and Institutional Economics Review*, 17.
- Wessels, W. J. (2000). *Economics*. Hauppauge: Barron's Educational Series, Inc.

Education

► Data Mining

Education and Training

Stephen T. Schroth
Department of Early Childhood Education,
Towson University, Baltimore, MD, USA

The use of big data, which involves the capture, collection, storage, collation, search, sharing, analysis, and visualization of enormous data sets so that this information may be used to spot trends, prevent problems, and to proactively engage in activities that make success more likely, has become increasingly popular and common. As the trend toward using big data has coincided with large-scale school reform efforts, which have provided increased data regarding student and teacher performance, operations, and the needs

of educational organizations, more and more school districts have turned to using big data to solve some of the problems they face. While certain leaders of schools and other organizations responsible for training students have rushed to embrace the use of big data, those concerned with student privacy have sometimes been critical of these attempts. The economic demands of setting up systems that permit the use of big data have also hindered some efforts by schools and training organizations to use this, as these bodies lack the infrastructure necessary to proceed with such efforts. As equipment and privacy concerns are overcome, however, the use of big data by schools, colleges, universities, and other training organizations seems likely to increase.

Background

Government agencies, businesses, colleges, universities, schools, hospitals, research centers, and a variety of other organizations have long collected data regarding their operations, clients, students, patients, and findings. With the emergence of computers and other electronic forms of data storage, more data than ever before began to be collected during the last two decades of the twentieth century. Because this data was often kept in separate databases, however, and was inaccessible to most users, much of the information that could be gleaned from it was not used. As technologies developed, however, many businesses became increasingly interested in making use of this information. Big data became seen as a way of organizing and using the numerous sources of information in ways that could benefit organizations and individuals.

By the late 1990s, interest in the field that became known as infonomics surged as companies and organizations wanted to make better use of the information they possessed, and to utilize it in ways that increased profitability. A variety of consulting firms and other organizations began working with large corporations and organizations in an effort to accomplish this. They defined *big data* as consisting of three “v”s, volume, variety, and velocity. *Volume*, as used in this context,

refers to the increase in data volume caused by technological innovation. This includes transaction-based data that has been gathered by corporations and organizations over time, but also includes unstructured data that derives from social media and other sources as well as increasing amounts of sensor and machine-to-machine data. For years, excessive data volume was a storage issue, as the cost of keeping much of this information was prohibitive. As storage costs have decreased, however, cost has diminished as a concern. Today, how best to determine relevance within large volumes of data, and how best to analyze data to create value have emerged as the primary issues facing those wishing to use it.

Velocity refers to the amount of data streaming in at great speed raises the issue of how best to deal with this in an appropriate way. Technological developments, such as sensors and smart meters, and client and patient needs, emphasize the necessity of overseeing and handling inundations of data in near-real time. Responding to data velocity in a timely manner represents an ongoing struggle for most corporations and other organizations. *Variety* in the types of formats in which data today comes to organizations presents a problem for many. Data today includes that in structured numeric forms which is stored in traditional databases, but has grown to include information created from business applications, e-mails, text documents, audio, video, financial transactions, and a host of others. Many corporations and organizations struggle with governing, managing, and merging different forms of data.

Some have added two additional criteria to these: variability and complexity. *Variability* concerns the potential inconsistency that data can demonstrate at times, which can be problematic for those who analyze the data. Variability can hamper the process of managing and handling the data. *Complexity* refers the intricate process that data management involves, in particular when large volumes of data come from multiple and disparate sources. For analysts and other users to fully understand the information that is contained in these data, they must be must first be connected, correlated, and linked in a way that helps users make sense of them.

Schools, colleges, universities, and other training centers have long had access to tremendous amounts of data concerning their students, teachers, and other operations. Demographic information, concerning age, gender, ethnicity, race, home language, addresses, parents' occupations, and other such data are collected as a matter of course. Evidence of students' academic performance also exists from a variety of sources, including teacher gradebooks, achievement tests, standardized tests, IQ tests, interest inventories, and a variety of other sources of information. As technological innovations such as computers, tablets, and other devices have become common in educational settings, it has become possible to gather an enormous amount of data related to how students think and perform, as well as how they make errors. As interest in school reform and improvement grew, so too did notice that a vast amount of data existed in education and training programs that was going unused. As a result, a great deal of effort has been put into attempts to create ways to harness this information through the use of big data analysis to offer solutions that might improve student performance.

Educational Applications

Educational and training programs have long collected data regarding students. Traditionally, however, much of this data remained in individual classrooms and schools, and was inaccessible by administrators and policy makers concerned with student learning. Although many local education authorities in the United States traditionally collected certain data regarding student performance, the federal *No Child Left Behind* legislation, passed in 2001, commenced a period when data regarding student performance in literacy and mathematics was collected to a greater degree than ever before. This practice was duplicated in most other nations, which resulted in an influx of data related to schools, teachers, and students. While much of this data was collected and transferred using traditional methods, over the past decade schools began using cloud storage that permitted easier access for district leaders.

Schools also started sending more data to state education agencies, which allowed it to be collected and analyzed in more sophisticated ways than ever before. As schools have increasingly used more programs, apps, tablets, and other electronic devices in an attempt to improve student performance, the amount of data has also grown. Schools and other organizations can now collect information that reflects not just student performance, but that indicates how a student thought about a problem when answering. Data can include individual keystrokes and deletions, eye movement, or how long a student held a mouse pointer above a certain answer.

Big data has been touted as providing many potential benefits for educational institutions and students. By providing the tools to collect and analyze data that schools, colleges, universities, and training programs already collect, big data will allow these educational institutions access to a series of predictive tools. These predictive tools will identify individual students' strengths and areas of need. As a result, educational and training programs will be able to improve learning outcomes for individual students by tailoring educational programs to these strengths and needs. A curriculum that collects data at each step of a student's learning process will permit schools, colleges, universities, and other training programs to meet student need on a daily basis. Educational and training programs will be able to offer differentiated assignments, feedback, units, and educational experiences that will promote optimal and more efficient learning experiences.

With this tremendous promise, big data's implementation and use is hindered by the needs for highly sophisticated hardware and software to permit real-time analysis of data. Using big data for the improvement of education and training programs requires massively parallel-processing (MPP) databases, which also require the ability to store and manage huge amounts of data. Search-based applications, data-mining processes, distributed file systems, the Internet, and cloud-based computing and storage resources and applications are also necessary. As most schools,

colleges, universities, and other training institutions lack a unified system, it has proven impossible for institutions to share such data on an internal basis, let alone across institutions. Unless and until these issues are resolved, big data will not have the capacity to permit all students to reach their full potential.

Privacy Issues and Other Concerns

Although using big data to permit students in schools, colleges, universities, and training programs has been trumpeted by many, including the United States Department of Education, many have objected to the process as endangering student privacy rights. Indeed, many schools, colleges, and universities lack rules, procedures, or policies that guide teachers and administrators regarding how much data to collect, how long to keep it, and who to permit access to it. Further, many schools, colleges, universities, and training programs have found themselves to be inundated by data, with little idea how best to respond. In an effort to best deal with this problem, many educational and training programs have sought to establish systems that would permit them to effectively deal with this.

In order to effectively use big data practices, schools, colleges, universities, and training programs must set up systems that permit them to store, process, and provide access to the data they collect. And as the data has grown to include not just student grades, but also attendance records, disciplinary actions, participation in sports, special education services provided, medical records, test performance, and the like. The data needs to be stored in a single database, in compatible formats, and accessible with a single password for the data to be used effectively. This infrastructure requires funding, and often the use of consultants or collaboration with other organizations.

As systems to accumulate and analyze data were established, many critics expressed fears that doing this might invade students' privacy rights, harm those who struggle, and allow data to fall into the hands of others. Many parents are

concerned, for example, that their child's early struggles with reading or mathematics could imperil their chances to be admitted to college, be bullied by peers, or looked at negatively by future employers. Fears have also been expressed that student data will be sold to commercial concerns. As the data held by schools becomes more comprehensive and varied, student disabilities, infractions, and other information that individuals might not want released is increasingly protected by those whom it concerns. This attitude has imperiled many attempts to use big data in educational and training settings.

Efforts to establish state-of-the-art systems to use big data procedures with students have met with opposition. In the United States, for example, the Bill & Melinda Gates Foundation and the not-for-profit Carnegie Corporation provided over \$100 million in funding for inBloom, a nonprofit organization that could provide the necessary technological support to permit K-12 schools to glean the benefits of big data. Although the states of Illinois, Massachusetts, and New York joined the process, the project was shut down after 2 years, largely because of opposition from parents and other privacy advocates. Despite this failure, other for-profit enterprises have been able to accumulate data from large numbers of students through programs that are sold to schools, who in turn receive information about student learning. Renaissance Learning, for example, sells the popular Accelerated Reader program that monitors students' reading comprehension to a global system of schools. As a result, it has accumulated data on over ten million students, and provides this to teachers and administrators who can use it to improve student performance.

Cross-References

- [Big Data Quality](#)
- [Correlation Versus Causation](#)
- [Curriculum, Higher Education, and Social Sciences](#)
- [Education](#)

Further Reading

- Foreman, J. W. (2014). *Data smart: Using data science to transform information into insight*. Hoboken: Wiley.
- Lane, J. E., & Zimpher, N. L. (2014). *Building a smarter university: Big data, innovation, and analytics*. Albany: The State University of New York Press.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data*. New York: Mariner Books.
- Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Hoboken: Wiley.

Electronic Commerce

- [E-Commerce](#)

Electronic Health Records (EHR)

Barbara Cook Overton
 Communication Studies, Louisiana State
 University, Baton Rouge, LA, USA
 Communication Studies, Southeastern Louisiana
 University, Hammond, LA, USA

Federal legislation required healthcare providers in the United States to adopt electronic health records (EHR) by 2015; however, transitioning from paper-based to electronic health records has been challenging. Obstacles include difficult-to-use systems, interoperability concerns, and the potential for EHRs negatively impacting provider-patient relationships. EHRs do offer some advantages, such as the ability to leverage data for insights into disease distribution and prevention, but those capabilities are underutilized. EHRs generate big data, but how to convert unstructured derivatives of patient care into useful and searchable information remains problematic.

EHRs were not widely used in American hospitals before the Health Information Technology for Economic and Clinical Health Act (HITECH) was passed by congress in 2009. HITECH required

hospitals receiving Medicaid and Medicare reimbursement to adopt and meaningfully use EHRs by 2015. The legislation was partly a response to reports released by the National Academy of Medicine (then called the Institute of Medicine) and the World Health Organization which, collectively, painted an abysmal picture of the American healthcare system. Respectively, the reports noted that medical mistakes were the eighth leading cause of patient deaths in the United States and that poor utilization of health information technologies contributed significantly to the US health system's low ranking in overall performance (the United States was ranked 37th in the world).

Public health agencies argued that medical errors could be reduced with the development and widespread use of health information technologies, such as EHRs. Studies suggested EHRs could both reduce medication errors and cut healthcare costs. It was predicted that improved access to patients' complete medical histories would help healthcare providers avoid duplicating treatment and overprescribing medications, thereby reducing medical errors, curtailing patient deaths, and saving billions of dollars. Despite the potential for improved patient safety and operational efficiency, pre-HITECH adoption rates were low because EHRs were expensive, difficult to use, and negatively affected provider-patient relationships. Evidence that EHRs would improve the quality of health care was neither conclusive nor straightforward. Nonetheless, the HITECH Act required hospitals to start using EHRs by 2015.

HITECH's major goals include reducing healthcare costs, improving quality of care, reducing medical errors, improving health information technology infrastructure through incentives and grant programs, and creating a national electronic health information exchange. Before HITECH was passed, only 10% of US hospitals used EHRs. By 2017, about 80% had some form of electronic charting. The increase is attributed to HITECH's meaningful use (MU) initiative, which is overseen by the Centers for Medicaid and Medicare Services. MU facilitated EHR adoption in two ways. First, it offered financial incentives for hospitals adopting and meaningfully using EHRs before the 2015 deadline ("meaningful use" is

defined as use that improves patient care, reduces disparities, and advances public health). Second, it imposed financial penalties for hospitals that failed to meet certain MU objectives by 2015 (penalties included withheld and/or delayed Medicare and Medicaid reimbursement).

Many MU requirements, however, are difficult to meet, costly to implement, and negatively impact provider productivity. Nearly 20% of early MU participants dropped out of the program, despite financial incentives and looming penalties. A majority of early MU participants – namely physicians – concluded that EHRs were not worth the cost, did not improve patient care, and did not facilitate coordination among providers. A survey of 1,000 physicians administered in 2013 revealed that nearly half believed EHRs made patient care worse and two-thirds reported significant financial losses following their EHR adoptions. Five years later, a Stanford Medicine poll found that 71% of physicians surveyed believed EHRs contributed to burnout and 59% thought EHRs needed a complete overhaul.

According to many healthcare providers, there are two main reasons EHRs need to be overhauled. The first has to do with ease of use. Most EHR systems were designed with billing departments in mind, not end users. Thus, the typical EHR interface resembles an accounting spreadsheet, not a medical chart. Moreover, the medical community's consensus that EHRs are hard to use has been widely documented. Providers contend that EHRs will not be fully functional or user-friendly until providers themselves are part of the design process.

The second reason providers believe EHRs need an overhaul centers on interoperability. Following HITECH passage in 2009, EHR makers rushed to meet the newly legislated demand. The result was dozens of proprietary software packages that did not talk to one another. This is especially problematic given HITECH's goals include standardized and interoperable EHRs. This means providers should be able to access and update health records even if patients seek treatment at multiple locations, but, as of 2020, most EHR systems were *not* interoperable. Consequently, the most difficult MU objective to meet,

according to several reports, is data exchange between providers.

Another factor complicating widespread EHR adoption is the widely held belief that EHRs negatively impact provider-patient relationships. Several studies show EHRs decrease the amount of interpersonal contact between providers and patients. For example, computers in exam rooms hinder communication between primary care providers and their patients: a third of the average physician's time is spent looking at the computer screen instead of the patient, and the physician, as a result, misses many of the patient's nonverbal cues. Other studies note that physicians' exam room use of diagnostic support tools, a common EHR feature, erodes patient confidence. For this reason, the American Medical Association urges physicians to complete as much data entry outside the exam room as possible. Studies also find many nurses, even when portable EHR workstations are available, opt to leave them outside of patients' rooms because of perceptions that EHRs interfere with nurse-patient relationships.

When compared with physicians, nurses have generally been more accepting of and enthusiastic about EHRs. Studies find more nurses than physicians claim EHRs are easy to use and help them complete documentation tasks more quickly. Nurses, compared with physicians, are considerably more likely to conclude that EHRs make their jobs easier. Despite concerns that EHRs can dehumanize healthcare delivery, nurses' positive attitudes are often rooted in their belief that EHRs improve patient safety.

EHRs are supposed to help keep patients safe by reducing the likelihood of medical mistakes occurring, but research finds EHRs have introduced new types of clinical errors. For example, "wrong-patient errors," which were infrequent when physicians used paper-based medical charts, are increasingly commonplace – physicians using EMRs regularly "misclick," thereby ordering, erroneously, medications and/or medical tests for the wrong patients. During an experiment, researchers observed that 77% of physicians did not confirm patients' identities before ordering laboratory tests. The study's authors attributed

many of the errors to poorly designed and hard-to-use EHRs.

In addition to increasing the likelihood of wrong-patient errors occurring, EHRs can affect patients in other ways as well. For example, EHRs can alter patients' perceptions of their healthcare providers. This is important because patient satisfaction is associated positively with healthy outcomes. Difficult-to-use EHRs have been shown to decrease providers' productivity and, thereby, increase patients' wait times and lengths of stay – two factors tied directly to patients feeling dissatisfied.

Patient satisfaction hinges on several factors, but one important determinant is whether and how patients tell their stories. Storytelling is a way for patients to make sense of uncertain circumstances, and patients who are allowed to tell their stories are generally more satisfied with their providers and typically have better outcomes. Patient narratives can also mean fewer diagnostic tests and lower healthcare costs. However, EHRs limit the amount of free text available for capturing patients' stories. Providers, instead, reduce narratives to actionable lists by checking boxes that correlate with patients' complaints and symptoms. Check boxes spread across multiple screens remove spontaneity from discourse, forcing patients to recite their medical histories, ailments, and medications in a prescribed fashion. Such medical records, comprised largely of numbers and test results, lack context.

EHRs generate and store tremendous amounts of data, which, like most big data, are text-heavy and unstructured. Unstructured data are not organized in meaningful ways, thereby restricting easy access and/or analysis. Structured data, by contrast, are well organized, searchable, and interoperable. Some EHR systems do code patient data so as to make the data partially structured, but increasing volumes of unstructured patient data handicap many healthcare systems. This is due, in large part, to hybrid paper-electronic systems. Before and during an EHR adoption, patients' health data are recorded in paper charts. Along with printouts of lab reports and digital imaging results, the medical chart is neither unified nor searchable. Scanning paper-

based items digitizes the bulk of the medical record, but most scanned items are not searchable via text recognition software. Handwritten notes, frequently copied and then scanned, are often illegible, further limiting access and usability. As documentation shifts from paper-based to electronic charting, more of patients' records become searchable. Although EHRs are not maximally optimized yet, they do present a clear advantage over paper-based systems: reams of patient data culled from medical and pharmaceutical records are no more searchable in paper form than unstructured big data. EHRs do offer a solution; however, leveraging that data requires skill and analytical tools.

Although an obvious benefit of using an EHR is readily accessible medical records, providers who lack the expertise and time necessary for searching and reviewing patients' histories often underutilize this feature. Evidence suggests some providers avoid searching EHRs for patients' histories. Instead, providers rely on their own memories or ask patients about previous visits. One study found that although physicians believed reviewing patients' medical records before examining them was important, less than a third did so. Thirty-five percent of physicians admitted that asking patients about their past visits was easier than using the EHR, and among those who tried using the EHR, 37% gave up because the task was too time-consuming.

A burgeoning field of health data analytics is poised to facilitate access and usability of EHR data. Healthcare analytics can help reduce costs while enhancing data exchange, care coordination, and overall health outcomes. This merger of medicine, statistics, and computer science can facilitate creating longitudinal records for patients seeking care in numerous venues and settings. This sets the stage for improved patient-centered health care and predictive medicine. Analytic tools can identify patients at risk for developing chronic conditions like diabetes, high blood pressure, or heart disease. Combining health records with behavioral data can enable population-wide predictions of disease occurrence and facilitate better prevention programs and improve public health.

EHRs, like paper medical charts, must be safeguarded against privacy and security threats. Patient privacy laws require encrypted data be stored on servers which are firewall- and password-protected. These measures afford improved control and protection of electronic health data, considering paper charts can be accessed, copied, or stolen by anyone entering a room where records are kept. Controlling access to electronic health data is accomplished by requiring usernames and passwords. Most EHRs also restrict access to certain portions of patients' data depending on users' level of authorization. For instance, while nurses may view physicians' progress notes and medication orders, they may not change them. Likewise, physicians cannot change nurses' notes. Techs may see which tests have been ordered, but not the results. This is important given EHRs are accessible by many providers, all of whom can contribute to the patient record.

EHRs, while promising, are not widely utilized nor efficiently leveraged for maximum productivity. Many are calling for a "next-generation" EHR prioritizing interoperability, information sharing, usability, and easily accessible/searchable data. Nonetheless, EHRs in their current form ensure data are largely protected from security breaches and are backed up regularly – these are clear advantages over paper charts susceptible to violation, theft, or damage (i.e., consider the tens of thousands of paper-based medical records destroyed by Hurricane Katrina). How EHRs affect provider productivity and provider-patient relationships are highly contested subjects, but evidence suggests enhanced data-mining capabilities can improve disease prevention and intervention efforts thereby improving health outcomes.

Cross-References

- ▶ [Biomedical Data](#)
- ▶ [Epidemiology](#)
- ▶ [Health Care Delivery](#)
- ▶ [Health Informatics](#)
- ▶ [Patient-Centered \(Personalized\) Health](#)
- ▶ [Patient Records](#)

Further Reading

- Adler-Milstein, J., et al. (2017). Electronic health record adoption in US hospitals: The emergence of a digital ‘advanced use’ divide. *Journal of the American Medical Informatics Association*, 24(6), 1142–1148.
- Christensen, T., & Grimsmo, A. (2008). Instant availability of patient records, but diminished availability of patient information: A multi-method study of GP’s use of electronic health records. *BMC Medical Informatics and Decision Making*, 8(12), 1–9.
- DesRoches, C. (2013). Meeting meaningful use criteria and managing patient populations: A National Survey of practicing physicians. *Annals of Internal Medicine*, 158, 791–799.
- Henneman, P., et al. (2008). Providers do not verify patient identity during computer order entry. *Academic Emergency Medicine*, 15(7), 641–648.
- Institute of Medicine. (1999). *To err is human: Building a safer health system*. Washington DC: National Academies Press.
- Montague, E., & Asan, O. (2014). Dynamic modeling of patient and physician eye gaze to understand the effects of electronic health records on doctor-patient communication and attention. *International Journal of Medical Informatics*, 83, 225–234.
- Nambisan, P., et al. (2013). Understanding electronic medical record adoption in the United States: Communication and sociocultural perspectives. *Interactive Journal of Medical Research*, 2, e5.
- Overton, B. (2020). *Unintended consequences of electronic medical records: An emergency room ethnography*. Lanham: Lexington Books.
- Stanford Medicine. (2018). *How doctors feel about electronic health records: National physician Poll by the Harris Poll*. Retrieved from <https://med.stanford.edu/content/dam/sm/ehr/documents/EHR-Poll-Presentation.pdf>.
- Stark, P. (2010). Congressional intent for the HITECH act. *The American Journal of Managed Care*, 16, SP24–SP28.

Ensemble Methods

Patrick Juola
Department of Mathematics and Computer
Science, McAnulty College and Graduate School
of Liberal Arts, Duquesne University, Pittsburgh,
PA, USA

Synonyms

Consensus methods; Mixture-of-experts

Ensemble methods are defined as “learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions” (Dietterich 2000). Assuming reasonable performance and diversity on the part of each of the component classifiers (Dietterich 2000), the collective answer should be more accurate than any individual member of the ensemble. For example, if the first classifier makes an error, the second and third classifiers, if correct, can “outvote” the first classifier and lead to a correct analysis of the overall system. Ensemble methods thus provide a simple and commonly used method of boosting performance in big data analytics.

Ensemble methods provide many advantages in big data classification. First, because the overall performance is generally better than that of any individual classifier in the ensemble, “you can often get away with using much simpler learners and still achieve great performers” (Gutierrez and Alton 2014). Second, these classifiers can often be trained in parallel on smaller subsets of data requiring less time and data access than a more sophisticated system that requires access to the entire dataset at once.

There are several different methods that can be used to construct ensembles. One of the easiest and most common methods is “bagging” (Breiman 1996; Dietterich 2000), where each classifier is trained on a randomly chosen subset of the original data. Each classifier is then given one vote and the overall prediction of the ensemble is the answer that receives the most votes. Other methods used include weighting votes by the measured accuracy of each classifier (a more accurate classifier receives greater weight), separating the training set into disjoint sets and cross-validating, or calculating probabilities and using Bayesian statistics to directly assess the probability of each answer. More esoteric methods may involve learning classifiers as well as learning additional selection algorithms to choose the best classifier or classifiers for any specific data point.

Another method of constructing ensembles is to use adaptive training sets in a procedure called “boosting.” Any specific classifier is likely to

perform better on some types of input than on others. If these areas of high and low performance can be identified, the boosting algorithm constructs a new training set that focuses on the mistakes of that classifier and trains a second classifier to deal with them. “Briefly, boosting works by training a set of learners sequentially and combining them for prediction, where the later learners focus more on the mistakes of the earlier learners” (Zhou 2012).

Within this framework, almost any learning or classification algorithm can be used to construct the individual classifiers. Some commonly used methods include linear discriminant analysis, decision trees, neural networks (including deep learning), naïve Bayes classifiers, k -nearest neighbor classifiers, and support vector machines. Other applications of ensemble methods include not only classification into categories, but also prediction of numeric values or discovering the structure of the data space via clustering.

Applications of ensemble methods include network intrusion detection, molecular bioactivity and protein locale prediction, pulmonary embolisms detection, customer relationship management, educational data mining, music and movie recommendations, object detection, and face recognition (Zhou 2012). Ensemble methods provide a powerful and easy to understand method of analyzing data that is too complicated for manual analysis.

Further Reading

- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems. MCS 2000* (Lecture notes in computer science) (Vol. 1857). Berlin/Heidelberg: Springer. https://doi.org/10.1007/3-540-45014-9_1.
- Gutierrez, D., & Alton, M. (2014). Ask a data scientist: Ensemble methods. *InsideBigData.com*. <https://insidebigdata.com/2014/12/18/ask-data-scientist-ensemble-methods/>.
- Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. Boca Raton: CRC Press.

Entertainment

Matthew Pittman and Kim Sheehan
School of Journalism & Communication,
University of Oregon, Eugene, OR, USA

Advances in digital technology have given most mobile devices the capability to not only stream but actually recognize (Shazam, VideoSurf, etc.) entertaining content. Streaming data is replacing rentals (for video) and hard disc ownership (for video and audio). Consumers have more platforms to watch entertaining content, more devices on which to watch it, and more ways to seek out new content. The flip side is that content producers have more ways to monitor and monetize who is consuming it. To complicate matters, user-generated content (YouTube videos, remixes, and social media activity) and metadata (data about data, or the tracking information attached to most files) are changing the need for – and enforcement of – copyright laws.

The traditional Hollywood distribution model (theatrical release, pay-per-view, rental, premium cable, commercial cable) has changed dramatically in the wake of smartphones, tablets, and similarly mobile devices on which people can now view movies and television shows. With DVD and Blu-ray sales declining, production studios are constantly experimenting with how soon to allow consumers to buy, rent, or stream a film after its theatrical release. Third-party platforms like Apple TV, Netflix, or Chromecast are competing with cable providers (Comcast, Time Warner, etc.) to be consumers’ method of choice for entertainment in and out of the home.

Andrew Wallenstein has said that, for content producers, the advent of big data will be like going from sipping through a straw to sucking on a fire hose: once they figure out how to wrangle all this new information, they will understand their customers with an unprecedented level of sophistication. Already, companies have learned to track everything users click on, watch, or scroll past in

order to target them with specific advertising, programming, and products. Netflix in particular is very proud of the algorithms behind their recommendation systems: they estimate that 75% of viewer activity is driven by their recommendations. Even when customers claim to watch foreign films or documentaries, Netflix knows more than enough to recommend programs based on what people are *actually* watching.

When it comes to entertainment, Netflix and Hulu also developed such successful models for distributing content that they have begun to create their own. Creating original content (in the case of Netflix, award-winning content) has solidified the status of Netflix, Amazon, Hulu, and others as the new source for entertainment content distribution.

On the consumer end, the migration of entertainment from traditional cable televisions to digital databases makes it difficult to track a program's ratings. In the past, ratings systems (like Neilson) used viewer diaries and television set meters to calculate audience size and demographic composition for various television shows. They knew which TV sets were tuned to what channels at what time slot, and this data was enormously useful for networks in figuring out programming schedules.

Online viewing, however, presents new challenges. There are enormous amounts of data that corporations have access to when someone watches their show online. Depending on the customer's privacy settings and browser, he or she might yield the following information about himself or herself to the site from which they are streaming audio or video: what other pages the customer has recently visited and in what order, his or her social media profile and thus demographic information, purchase history and items which he or she might currently be in the market, or what other media programs he or she watches. New algorithms are constantly threatening the delicate balance between privacy and convenience.

With traditional ratings, the measurements occurred in real time. However, most online viewing (including DVR- or TiVo-mediated viewing)

occurs whenever it is convenient for the customer, which may be hours, days, weeks, or even years after the content originally aired. The issue then becomes how long after entertainment is posted to count a single viewing toward its ratings. Shows like *Community* might fail to earn high enough ratings to stay with NBC, the network that originally produced it. NBC would initially air an episode via traditional network broadcast and then host it online the next day. However, thanks to *Community*'s loyal following online, it found a new digital home: the show is now produced and hosted online by Yahoo TV. As long as there is a fan base for a kind of entertainment, its producers and consumers should always be able to find each other amid the sea of digital data.

Musical entertainment is undergoing a similar shift in the age of big data. Record companies now license their music to various platforms: the iTunes store came out in 2003, Pandora in 2004, and Spotify in 2006. These digital music services let users buy songs, listen to radio, and stream songs, respectively. Like with Netflix and its videos, music algorithms have been developed to help consumers find new artists with a similar sound to a familiar one. Also like with video, the digital data stream of consumption lets companies know who is listening to their product, when they listen, and through what device.

Analytics are increasingly important for producers and consumers. The band Iron Maiden found that lots of people in South America were illegally downloading their music, so they put on a single concert in São Paulo and made \$2.58 million. Netflix initially paid users to watch and metatag videos and came up with over 76,000 unique ways to describe types of movies. Combining these tags with customer viewing habits led to shows people actually wanted: *House of Cards* and *Orange Is the New Black*. Hulu experimented with a feature that let users search for words in captions on a show's page. So while looking at, say, *Parks and Recreation*, if users frequently searched for "bloopers" or "Andy naked prank," Hulu could prioritize that content. The age of big data has wrought an almost limitless number of

ways to entertain, be entertained, and keep track of that entertainment.

Cross-References

► [Netflix](#)

Further Reading

- Barnes, S. B. (2006). A privacy paradox: Social networking in the United States. *First Monday*, 11(9), 0–14. http://firstmonday.org/article/view/1394/1312_2
- Breen, C. Why the iTunes store succeeded. <http://www.macworld.com/article/2036361/why-the-itunes-store-succeeded.html>. Accessed Sept 2014.
- Schlieski, T., & Johnson, B. D. (2012). Entertainment in the age of big data. *Proceedings of the IEEE*, 100 (Special Centennial Issue), 1404–1408.
- Vanderbilt, T. The science behind the Netflix algorithms that decide what you'll watch next. http://www.wired.com/2013/08/qq_netflix-algorithm/. Accessed Sept 2014.

Environment

Zerrin Savasan

Department of International Relations, Sub-
Department of International Law, Selçuk
University, Konya, Turkey

The environment phenomena include both “that environs” and “what is environed” and the relationship between the “environing” and the “environed.” It can be understood as all the physical and biological surroundings involving linkages/interrelationships at different scales between different elements. It can also be defined as all natural elements from ecosystem to biosphere plus human-based elements and their interactions. If a clear grasp of the term cannot be rendered as a first step, it cannot be understood correctly what is meant by the term in related subjects. Therefore, it can be applied wrongly/or incompletely in the processes of studying these subjects, due to the substantial divergences in the understanding of

the term. So, here, it is firstly required to clarify what is really said by the term environment.

The Term Environment

The term environment in fact can be defined in several different ways and can be used in various forms, contexts. To illustrate, its definition can be categorized in four basic categories and several subcategories: 1. building blocks, 1.1. architectural (built environment-natural environment), 1.2. geographical (terrestrial environment-aquatic environment), 1.3. institutional (home environment-work environment-social environment); 2. economic uses, 2.1. inputs (natural resources-system services), 2.2. outputs (contamination-products), 2.3. others (occupational health-environmental engineering); 3. spatial uses, 3.1. ecosystems (forest-rangeland-planet), 3.2. comprehensive (watershed-landscape); 4. ethical/spiritual uses, 4.1. home (nature-place-planet-earth), 4.2. spiritual (deep ecology-culture-wilderness-GAIA).

However, given the definitions of some dictionaries, it is generally defined as the external and internal conditions affecting the development and survival of an organism and ultimately giving its form; or the sum of social and cultural conditions influencing the existence and growth of an individual or community's life. As commonly used, the term environment is usually understood as the surrounding which an organism finds itself immersed in. It is actually a more complex term involving more than that, because it includes all aspects or elements that affect that organisms in distinct ways and each organism in turn affects all those which affect itself. That is, each organism is surrounded by all those influencing each other through a causal relationship.

Related Terms: Nature

In its narrow sense the environment implies all in nature from ecosystem to biosphere, on which there is no human impact or the human impact is kept under a limited level. Most probably because of that, for many, the terms environment and

nature have been used interchangeably, and it is so often thought that the term environment is synonymous with nature. Yet, the term nature consists of all on the earth, but not the human-made elements. So, while the word environment is used, it means more than the nature, so actually should not be substituted for the nature. In its more broadly usage, it refers to all in nature in which all human beings and all other living organisms, plants, animals, etc., have their being and interrelationships among all in nature and with the living organisms. That means, it covers natural aspects as well as human-made aspects (represented by built environment). Therefore, it can be classified into two primary dimensions.

1. **Natural (or Physical) Dimension:** It encompasses all living and nonliving things occurring naturally on earth, so, two components, abiotic (or nonliving) and biotic (or living) component.
 - **Abiotic (or nonliving)** component involves physical factors including sunlight (essential for photosynthesis), precipitation, temperature, and types of soil present (sandy or clay, dry or wet, fertile or infertile to ensure base and nutrients); and chemical factors containing proteins, carbohydrates, fats, and minerals. These elements establish the base for further studies on living components.
 - **Biotic (or living)** component comprises plants, animals, and microorganisms in complex communities. It can be distinguished by three types.
 - (a) **Producers/autotrophs:** The producers absorb some of the solar energy from the sun and transform it into nutritive energy through photosynthesis, i.e., they are self-nourishing organisms preparing organic compounds from inorganic raw materials through the processes of photosynthesis, e.g., all green plants, both terrestrial and aquatic ones such as phytoplankton.
 - (b) **Consumers/heterotrophs:** The consumers depend on the producers for energy directly (herbivores such as rabbits) or indirectly (carnivores such as tigers). When they consume the plants, they absorb their chemical energy into their bodies, and thus, make use of this energy in their bodies to maintain their livelihood, e.g., animals of all sizes ranging from large predators to small parasites, e.g., herbivores, carnivores, omnivores, mosquito, flies, etc.
 - (c) **Decomposers:** When plants and animals die, the rest of the chemical energy staying in the consumers' bodies are used by the decomposers. The decomposers convert the complex organic compounds of these dead plants and animals to simpler ones by the processes of decomposition and disintegration, e.g., microorganisms such as fungi, bacteria, yeast, etc., as well as a diversity of worms, insects, and many other small animals.
2. **Human-Based (or Cultural) Dimension:** It basically includes all human-driven characteristics of the environment, so its all components that are strongly influenced by human beings. While living in the natural environment, human beings change it to their needs, they accept norms, values, and make regulations, manage economic relations, find new technologies, establish institutions and administrative procedures, and form policies to conduct them, so in brief, create a new environment to meet their survival requirements modifying the natural environment.

Related Terms: Ecosystem/Ecology

Another term which is used often interchangeably with the environment is ecosystem. Like the term nature, ecosystem is also used as synonymously with the environment. This is particularly because the research subjects of all sciences related to the environment are interconnected and interrelated to each other. To illustrate, natural science is

concerned with the understanding of natural phenomena on the basis of observation and [empirical evidence](#). In addition, earth science which is one of the branches of natural science provides the studies of the atmosphere, hydrosphere, lithosphere, and biosphere. Ecology, on the other hand, as the scientific study of ecosystem, is defined as a discipline studying on the interactions between some type of organism and its nonliving environment and so on how the natural world works. In other words, its research area is basically restricted to the living (biotic) elements in the nature, i.e., the individual species of plants and animals or community patterns of interdependent organisms which along with their nonliving environment including the atmosphere, geosphere, and hydrosphere. Thus, ecology arises as a science working like the biological science of environmental studies.

Nevertheless, particularly after the increasing role of human component in both disciplines, i.e., in both environmental studies and ecology, the difference on the research subjects of two sciences has almost been eliminated. Hence, currently, both the environmental scientists and ecologists examine the impacts of linkages in the nature and also interactions and interrelationships of living (biotic) and nonliving (abiotic) elements with each other. Their investigations are so mostly rested on similar methods and approaches.

Based on these facts, the question here arises then what the concept ecosystem means, what should be understood from that concept as different from the concept of environment. Ecosystem can be simply identified as an interacting system in which the total array of plant and animal species (biological component) inhabiting a common area in relationship to their nonliving environment (physical component) having effects on one another interdependently. So, it constitutes a significant unit of the environment, and environmental studies. Accordingly, it should be underlined that even if two terms – environment and ecosystem – are both deeply interrelated terms concerned with nature and their scientific studies using the similar perspectives, they should not be substituted for each other. This is particularly because the two terms differ dramatically in their

definitions. An environment is established by the surroundings (involving both natural environment and human-based environment) in which we live in; an ecosystem is a community of organisms (biotic) functioning with an environment. In order to make it easier to understand, it is usually studied as divided into two major categories: aquatic (or water) ecosystem, such as lakes, seas, streams, rivers, ponds etc.; terrestrial (or land) ecosystem, such as deserts, forests, grasslands, etc. However, in fact, while a lake or a forest can be considered as an ecosystem, the whole structure of the earth involving interrelated set of smaller systems also forms an ecosystem, referred to as ecosphere (or global ecosystem), the ecosystem of earth.

Human Impact on Environment

The humankind is dependent on the environment for their survival, well-being, continued growth and so their evolution and the environment is dependent on the humankind for its conservation and evolution. So, there is an obvious interdependent relationship between humankind development and their environment. The humans should be aware of the fact that while they are degrading the environment, they are actually harming themselves and preparing their own end. Yet, unfortunately, till date, the general attitude of human beings has been to focus on their development rather than the protection and development of the environment. Indeed, enormous increase in human population has required new needs in greater numbers and so raised the demand for constantly increasing development. It has resulted in growing excesses of industrialization and technology development to facilitate transformation of resources rapidly into the needs of humans, and so increasing consumption of various natural resources. Various sources of environmental pollution (air, land, water), deforestation and climate change which are among the most threatening environmental problems today, have also generated in human activities. Therefore, it is generally admitted that human intervention has been a very crucial factor changing the

environment, although it is sometimes positively, unfortunately often negatively, causing large scale environmental degradation.

There are various environments as understood from above mentioned explanations ranging from those at very small scales to the entire environment itself. They are all closely linked to each other, so the presence of adverse effects in even small-scale environment may ultimately be followed by environmental degradation at entire world. The human being realizes this fact and environmental problems have started to be seen as a major cause of global concern with late 1960s. Since then, a great number of massive efforts – agreements, organizations, and mechanisms – have been created and developed to create awareness about environment pollution and related adverse effects, and about the humans' responsibilities towards the environment, and also to form meaningful support towards environmental protection. They all, working in the fields concerning global environmental protection, have been initiated to reduce these problems which require to be coped with through a globally concerted environmental policy.

Particularly, the United Nations (UN) system, developing and improving international environmental law and environmental policy, with its crucial organs, significant global conferences like Stockholm Conference establishing the United Nations Environment Programme (UNEP), Rio Conference establishing the Commission on Sustainable Development (CSD), and numerous specialized agencies such as International Labour Organization (ILO), World Health Organization (WHO), International Monetary Fund (IMF), United Nations Educational, Scientific and Cultural Organization (UNESCO), semi-autonomous bodies like the UN Development Programme (UNDP), UN Institute for Training and Research (UNITAR), the UN Conference on Trade and Development (UNCTAD), and the UN Industrial Development Organization (UNIDO), has greatly contributed to the struggle with the challenges of global environmental issues. Moreover, since the first multilateral environmental agreement (MEA), Convention on the Rhine, adopted in 1868, the number of MEAs have

gone up rapidly, particularly in the period from Stockholm to Rio.

The Goal of Sustainable Development

Recently, in the Rio+20 United Nations Conference on Sustainable Development (UNCSD), held in Rio de Janeiro, Brazil, from 20 to 22 June 2012, while the seriousness of global environmental deterioration is acknowledged, at the same time the importance of the goal of sustainable development as a priority is re-emphasized. Indeed, its basic themes are building a green economy for sustainable development including support for developing countries and also building an institutional framework for sustainable development to improve international coordination. The renewed political commitment for sustainable development – implying simply the integration of the environment and development, and more elaborately, development meeting the needs of the present without compromising the ability of future generations to meet their own needs, as defined in Brundtland Report (1987) prepared by World Commission on Environment and Development – is also reaffirmed by the document, namely “The Future We Want,” created by the Conference. This document also supports on the development of 17 measurable targets aimed at promoting sustainable development globally, namely, [Sustainable Development Goals](#) (SDGs) of the 2030 Agenda for Sustainable Development. These goals adopted by the UN Headquarters held in September 2017 include the followings: ending poverty and hunger, ensuring healthy lives, inclusive and equitable quality education, gender equality, clean water and sanitation, affordable and clean energy, decent economy, sustainable industrialization, sustainable cities and communities, responsible consumption and production, climate action, peace, justice and strong institutions, partnerships for the goals, and reducing inequalities. They are built on the eight Millennium Development Goals (MDGs) adopted in September 2000 at another UN Headquarters, setting out a series of targets such as eradicating extreme poverty/hunger, improving universal

primary education, gender equality, maternal health, environmental sustainability, global partnership for development, reducing child mortality, and coping with HIV/AIDS, malaria, and other diseases with a deadline of 2015.

The foundation for the concept of sustainable development is laid firstly through the Founex report (Report on Development and Environment) which is prepared by a panel of experts meeting in Founex, Switzerland, in June 1971. Indeed, according to the Founex report, while the degradation of the environment in wealthy countries is mainly as a result of their development model, in developing countries it is a consequence of underdevelopment and poverty. Then, the official title of the UN Conference on Environment and Development (UNCED), held in Rio de Janeiro, in 1992 in itself summarizes, in fact, the efforts of the UN Conference on the Human Environment (UNCHE), held in Stockholm, in 1972, or rather, those of the Founex Report. The concept has been specifically popularized by the official titles of two last UN Conferences, namely, the World Summit on Sustainable Development, held in Johannesburg, in 2002 and the United Nations Conference on Sustainable Development (UNCSD), held in Rio de Janeiro, in 2012.

Intelligent Management of the Environment: Big Data

As mentioned above, the humankind has remarkably developed itself on learning the ways of reconciling both the environmental and developmental needs, and thus on achieving a sustainable relationship with the environment. Yet, despite all those efforts, it seems that comprehension and intelligent management of the environment is still inadequate and incomplete. It remains as one of the most important challenges of the humankind to face with. This is particularly because of two fundamental reasons.

1. Environment is a multidisciplinary subject encompassing diverse fields that should be examined from many different aspects. It should include the studies of different sciences

such as economics, geology, geography, hydrology, history, physics, physiology, etc.

2. Environmental science examining basically environment has a direct relevance to different sides of life of all living beings. It is a multi-disciplinary science involving many different research topics like protection of nature and natural resources, biological diversity, prevention and reduction of environmental pollution, stabilization of human population, the relation between development and environment, improvement of modern technologies supporting renewable energy systems etc.

This extraordinarily broad field stemming from both the environment's and environmental science's multiplicity results in an explosion in the data type/amount/methods of storage/models of usage, etc., at an unprecedented rate. To capture this complexity and diversity and to better understand the field of environment, to address the challenges associated to environmental problems/sustainable development (Keeso (2014) argues that while Big Data can become an integral element of environmental sustainability, Big Data can make environmental sustainability an essential part of its analysis vice versa, through emerging new tools such as collaborative partnerships and business model innovation), it is very recently suggested to build Big Data sets, having a multi-disciplinary dimension encompassing diverse fields in themselves and having influence within multiple disciplines.

Although there is no common understanding/identification on Big Data (ELI 2014; Keeso 2014; Simon 2013; Sowe and Zettsu 2014), it generally refers to large-scale and technology-driven/computer-based data collection/storage/analysis, in which data obtaining/monitoring/estimating is quick and easy by the means of satellites, sensor technology, and models. Therefore, it is often argued that, by means of Big Data, it becomes easier to identify vulnerabilities requiring further environmental protection and to make qualified estimations on future prospects, thus to take preventive measures urgently to respond to environmental problems and in the final analysis to reduce hazardous exposures of environmental

issues, ranging from climate change to air-land-water pollution.

In exploring how these large-scale and complex data sets are being used to cope with environmental problems in a more predictive/preventive and responsive manner, the following cases can be shown as examples (ELI 2014):

- Environmental Maps generated from US Environmental Protection Agency (EPA) databases including information on environmental activities the context of EnviroMapper
- Online accession to the state Departments of Natural Resources (DNRs) and other agencies for Geographic Information Systems (GIS) data on environmental concerns
- Usage of Big Data sets in many states and localities' environmental programs/in the administration of their federally delegated programs
- The Green Initiatives Tracking Tool (GITT) developed by the US Postal Service to collect information on employee-led sustainability projects – related to energy, water, and fuel consumption and waste generation – taking place across its individual facilities
- Collection of site-based data by the National Ecological Observatory Network (NEON) related to the effects of climate change, land use change, and invasive species from several sites throughout the USA
- The Tropical Ecology Assessment and Monitoring Network (TEAM) of publicly shared datasets developed by Conservation International (CI) to serve as an early warning system to alert about environmental concerns, to monitor the effects of climate or land use changes on natural resources and ecosystems
- Country/issue ranking on countries' management of environmental issues and investigation of global data comparing environmental performance with GDP, population, land area, or other variables by a Data Explorer under the context of the Environmental Performance Index (EPI).

As shown above by example cases, the use of Big Data technologies on environment-related

issues gradually increases, yet, still there is need for further research for tackling with challenges raised about the use of Big Data (Boyd 2010; Boyd and Crawford 2012; De Mauro et al. 2016; Forte Wares, —; Keeso 2014; Mayer-Schönberger and Cukier 2013; Simon 2013; Sowe and Zettsu 2014).

Cross-References

- Earth Science
- Pollution, Air
- Pollution, Land
- Pollution, Water

Further Reading

- Boyd, D. (2010). Privacy and publicity in the context of big data. WWW Conference, Raleigh, 29 Apr 2010. Retrieved from <http://www.danah.org/papers/talks/2010/WWW2010.html>. Accessed 3 Feb 2017.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662–679. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878>. Accessed 3 Feb 2017.
- De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. Retrieved from https://www.researchgate.net/publication/299379163_A_formal_definition_of_Big_Data_based_on_its_essential_features. Accessed 3 Feb 2017.
- Environmental Law Institute (ELI). (2014). *Big data and environmental protection: An initial survey of public and private initiatives*. Washington, DC: Environmental Law Institute. Retrieved from <https://www.eli.org/sites/default/files/eli-pubs/big-data-and-environmental-protection.pdf>. Accessed 3 Feb 2017.
- Environmental Performance Index (EPI)(-). Available at: <http://epi.yale.edu/>. Accessed 3 Feb 2017.
- Forte Wares(-). Failure to launch: From big data to big decisions why velocity, variety and volume is not improving decision making and how to fix it. White Paper. A Forte Consultancy Group Company. Retrieved from http://www.fortewares.com/Administrator/userfiles/Banner/forte-wares-pro-active-reporting_EN.pdf. Accessed 3 Feb 2017.
- Keeso, A. (2014). Big data and environmental sustainability: A conversation starter. Smith School Working Paper Series, Dec 2014, Working paper 14-04. Retrieved from <http://www.smithschool.ox.ac.uk/library/working-papers/workingpaper%2014-04.pdf>. Accessed 3 Feb 2017.

- Kemp, D. D. (2004). *Exploring environmental issues*. London/New York: Taylor and Francis.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London: John Murray.
- Patten, B. C. (1978). Systems approach to the concept of environment. *The Ohio Journal of Science*, 78(4), 206–222.
- Raven, P. H., & Berg, L. R. (2006). *Environment*. Danvers: John Wiley & Sons.
- Saunier, R. E., & Meganck, R. A. (2007). *Dictionary and Introduction to Global Environmental Governance*. London: Earthscan.
- Simon, P. (2013). *Too big to ignore: The business case for big data*. Hoboken: Wiley.
- Sowe, S. K., & Zettsu, K. (2014). Curating big data made simple: Perspectives from scientific communities. *Big Data*, 2(1), 23–33. Mary Ann Liebert, Inc.
- Withgott, J., & Brennan, S. (2011). *Environment*. New York: Pearson.

Epidemiology

David Brown^{1,2} and Stephen W. Brown³

¹Southern New Hampshire University, University of Central Florida College of Medicine, Huntington Beach, CA, USA

²University of Wyoming, Laramie, WY, USA

³Alliant International University, San Diego, CA, USA

Epidemiology is the scientific discipline concerned with the causes, the effects, the description of, and the quantification of health phenomena in specific identifiable populations. Epidemiologists, the public health professionals who study and apply epidemiology, investigate the geographic, the behavioral, the economic, the hereditary, and the lifestyle patterns that increase or decrease the likelihood of disease or injury in specific populations. The art and science of epidemiology investigates the worldwide outbreak of diseases and injury in different populations throughout the world. Epidemiological data is used to understand the distribution of disease and injury in an attempt to improve peoples' health and prevent future negative health consequences.

The primary goals of epidemiology are:

To *describe* the health status of populations and population subgroups. This information is used to develop statistical models showing how different groups of people are affected by different diseases and other health consequences. Big data in the form of demographic information is essential for the descriptive process of epidemiology.

To *explain* the etiological and causative factors that lead to or protect against disease or injury. Explanatory epidemiological data is also used to determine the ways in which disease and other health phenomena are transmitted. Big data are essential for the accurate identification of both causative factors and patterns of disease transmission.

To *predict* the occurrence of disease and the probability of outbreaks and epidemics. Predictive data are also used to estimate the positive effects of positive scientific and social changes such as the development of new vaccines and lifestyle changes such as increasing the amount of time people spend exercising. Big data allows for expanded data collection, improved data analysis, and increased information dissemination; these factors clearly improve prediction accuracy and timeliness.

To *control* the distribution and transmission of disease and other negative health events and to promote factors that improve health. The activities of describing, explaining, and predicting come together in implementing public health's most important function that improves national and world health. Epidemiological big data is used to identify potential curative factors for people who have contracted a disease. Big data identify factors that have the potential of preventing future outbreaks of disease and epidemics. Big data can also help identify areas where health education and health promotion activities are most needed and have the potential of having a positive impact.

Epidemiological research is divided into two broad and interrelated types of studies: descriptive epidemiological research and analytic

epidemiological research. Both of these types of studies have been greatly influenced by big data.

Descriptive epidemiology addresses such questions as: Who contracts and who does not contract some specific disease? What are the *people factors* (e.g., age, ethnicity, occupation, lifestyle, substance use and abuse) that affect the likelihood of contracting or not contracting some specific health problem? What are the *place factors* (e.g., continent, country, state, province, residence, work space, places visited) that affect the probability of contracting or not contracting some specific health problem? What are the “time factors” (e.g., time before diagnosis, time after exposure before symptoms occur) that affect the course of a health problem? Clearly, big data provides much needed information for the performance of all descriptive epidemiological tasks.

Analytic epidemiological studies typically test hypotheses about the relationships between specific behaviors (e.g., smoking cigarettes, eating a balanced diet) and exposures to specific events (e.g., experiencing a trauma, receiving an inheritance, being exposed to a disease) and the mortality (e.g., people in a population who die in a specific time period) and morbidity (e.g., people in a population who are ill in a specific time period). Analytic epidemiological studies require the use of a comparison group.

There are two types of analytic studies: prospective and retrospective. A prospective epidemiological study looks at the consequence of specific behaviors and specific exposures. As an example, a prospective analytic study might investigate the future lung cancer rate among people who currently smoke cigarettes with the future lung cancer rate of people who do not smoke cigarettes.

In a retrospective epidemiological study, the researcher identifies people who currently have a certain illness or other health condition, and then he or she identifies a comparable group of these people who do not have the illness or health condition. Then, the retrospective researcher uses many investigative techniques to look back in time to identify events or situations that

occurred in the past that differentiate between the two groups.

It is well documented that the four Vs of big data, volume, velocity, variety, and veracity, are well applied in the discipline of epidemiology. In examining the methods and the applications of epidemiology, it is apparent that the amount of information (volume) that is available via big data is a resource that will continue to help descriptive epidemiologists identify the people, place, and time factors of health and disease. The speed (velocity) at which big data can be collected, analyzed, disseminated, and accessed will continue to provide epidemiologists improved methods conducting analytic epidemiological studies. The different types of big data (variety) that are available for epidemiological analysis offer epidemiologists opportunities for research and application that were not even thought of only a few years ago. The truth (veracity) that big data provides the discipline of epidemiology can only lead to improvements in processes of health promotion and disease prevention.

Early Uses of Big Data in Epidemiology

The discipline of epidemiology can be traced back to Hippocrates who speculated about the relationship between environmental factors and the incidence of disease. Later, in 1663, an armature statistician, John Graunt, published a study concerning the effects of the bubonic plague on the mortality rates in London. Using these data, Graunt was able to produce the first life table that estimated the probabilities of survival for different age groups. However, the beginning of modern epidemiology, and a precursor to the use of what we now call big data, can be traced back to Dr. John Snow and his work with the cholera epidemic that broke out in London, England, in the mid-1800s.

The father of modern epidemiology, John Snow, was a physician who practiced in London, England. As a physician, Dr. Snow was very aware of the London epidemic of cholera that broke out in the mid-1800s. Cholera is a

potentially deadly bacterial intestinal infection that is caused by, and transmitted through, the ingestion of contaminated water or foods. Snow used what for his day were big data techniques and technologies. In doing so, Dr. Snow had a map of London on which he plotted the geographical location of those people who contracted cholera. By studying his disease distribution map, Snow was able to identify geographical areas that had the highest cluster of people who were infected with the disease. Through further investigation of the high outbreak area, Dr. Snow determined and demonstrated that the population in the affected area received their water from a specific source known as the Broad Street pump. Snow persuaded public officials to close the well that in turn led to a significant decrease in the incidence of cholera in that community. This mapping and plotting of the incidence of disease and the application of his discovery for the improvement of public health were the beginning of the discipline of epidemiology, and for its day, it was a clear and practical application of big data and big data techniques.

Contemporary and Future Uses of Big Data in Epidemiology

The value of epidemiological science is only as strong as the data it has at its disposal. In the early years and up until fairly recently, epidemiological data were collected from self-report from infected patients and reports from the practitioners who diagnosed the patients. These individuals had the responsibility for reporting the occurrence and characteristics of the problem to a designated reporting agency (e.g., the Centers for Disease Control and Prevention, a local health department, a state health department). The reporting agency then entered the data into a database that may have been difficult to share with other reporting agencies. As much as possible, the available data were shared with epidemiologists, statisticians, medical professionals, public health professionals, and health educators who would use the data to facilitate positive health outcomes.

Now in the age of big data, data collection, analysis, retrieval, and dissemination have greatly improved the previous process. Some of the ways big data is being used in epidemiology involve the use of the geographic information systems (GIS). Information from this technology is being used by health program planners and epidemiologists to identify and target specific public health interventions that will meet the needs of specific populations. GPS data has enabled expanded and improved methods for disease tracking and mapping. In many cases, smartphones enable patients with chronic diseases to transmit both location and symptom data that enables epidemiologists to find correlates between environmental factors and symptom exacerbation.

Big data application in public health and health care is currently being used in many other areas, and it is only reasonable to expect that over time these uses will expand and become more sophisticated. Some of the currently functioning big data applications include the development of a sophisticated international database to trace the clinical outcomes and cost of cancer and related disorders. In addition, medical scientists and public health specialists have developed a large international big data system to share information about spinal cord injury and its rehabilitation. Another big data application has been in the area of matching donors and recipients for organ transplantation.

Several major scientific and research groups have been convened to discuss big data and epidemiology. The consensus of opinion is that a major value of big data is dependent upon the willingness of scientists to share their data, their methodologies, and findings in open settings and that researchers need to work collaboratively on the same and similar problems. With such cooperation and team efforts, big data is seen as having great potential to improve the nation's and the world's health.

As more sophisticated data sets, statistical models, and software programs are developed, it is logical to predict that the epidemiological applications of big data will expand and become more sophisticated. As a corollary to this prediction, it is also reasonable to predict that big data will have

many significant contributions to world health and safety.

Conclusion

Epidemiology is a public health discipline that studies the causes, the effects, the description of, and the quantification of health phenomena in specific identifiable populations. Since the discipline concerns world populations, the field is a natural area for the application of big data. It should be noted that several scientific conferences have been conducted. The consensus from these conferences is that big data has the potential of making great strides in the improvement of world health epidemiological studies. However, they also note that in order to achieve this potential, data will need to be shared openly and that researchers will need to work cooperatively in their use of big data.

Cross-References

- [Biomedical Data](#)
- [Data Quality Management](#)
- [Prevention](#)

Further Reading

- Andrejevic, M., & Gates, K. (2014). Big data surveillance: Introduction. *Surveillance & Society*, 12(2), 185–196.
- Kao, R. R., Haydon, D. T., Lycett, S. J., & Murcia, P. R. (2014). Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends in Microbiology*, 22(5), 282–291.
- Marathe, M. V., & Ramakrishnan, N. (2013). Recent advances in computational epidemiology. *IEEE Intelligent Systems*, 28(4), 96–101.
- Massie, A. B., Kuricka, L. M., & Segev, D. L. (2014). Big data in organ transplantation: Registries and administrative claims. *American Journal of Transplantation*, 14(8), 1723–1730.
- Michael, K., & Miller, K. W. (2013). Big data: New opportunities and new challenges. *Computer*, 46(6), 22–24.
- Naimi, A. I., & Westreich, D. J. (2014). Big data: A revolution that will transform how we live, work, and think. *American Journal of Epidemiology*, 179(9), 1143.

Nielson, J. L., et al. (2014). Development of a database for translational spinal cord injury research. *Journal of Neurotrauma*, 31(21), 1789–1799.

Vachon, D. (2005). Doctor John Snow blames water pollution for cholera epidemic. *Old News*, 16(8), 8–10.

Webster, M., & Kumar, V. S. (2014). Big data diagnostics. *Clinical Chemistry*, 60(8), 1130–1132.

Error Tracing

- [Anomaly Detection](#)

Ethical and Legal Issues

Rochelle E. Tractenberg

Collaborative for Research on Outcomes and – Metrics, Washington, DC, USA

Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and

Rehabilitation Medicine, Georgetown University, Washington, DC, USA

Definition

“Ethical and legal issues” are a subset of “ethical, legal, and social issues” – or ELSI – where the “I” sometimes also refers to “implications.” This construct was introduced in 1989 by the National Program Advisory Committee on the Human Genome in the United States, with the intention of supporting exploration, discussion, and eventually the development of policies that anticipate and address the ethical, legal, and social implications of/issues arising from the advanced and speedily advancing technology associated with genome mapping. Since this time, and with ever-increasing technological advances that have the potential to adversely affect individuals and groups (much as genome research has) – including both research and non-research work that involves big data – ELSI relating to these domains are active areas of research and policy development. Since social implications of/issues arising from big data are discussed elsewhere in this

encyclopedia, this entry focuses on just ethical and legal implications and issues.

Introduction

There are ELSI in both research and nonscience analysis involving big data. However, whatever they are known to be right now, one of the principal issues in both research and other analysis of big data is actually that *we cannot know or even anticipate what they can be in the future*. Therefore, training in the identification of, and appropriate reaction to, ELSI is a universally acknowledged need; but this training must be comprehensive without being overly burdensome, and everyone involved should reject the lingering notion that either the identification or response to an ethical or legal problem “is just common sense.” That ethical professional practice that is simply a matter of common sense is an outdated – never correct – idea representing the perspective that scientists work within their own disciplinary silo in which all participants follow the same set of cultural norms. Modern work – with big data – involves multiple disciplines and is not uniquely/always scientific. The fact that these “cultural norms” have always been ideals, rather than standards to which all members of the scientific community are, or could be, held, is itself an ethical issue for research and analysis in big data. The effective communication of these ideals to *all* who will engage with big data, whether as researchers or nonscientific data analysts, is essential.

Legal Implications/Issues

Legal issues have traditionally been focused on liability (i.e., limiting this), e.g., for engineering and other disciplines where legal action can result from mistakes, errors, or malfeasance. In the scientific domains, legal issues tend to focus only on plagiarism, falsification of data or results, and fraud, e.g., making false claims in order to secure funding for research and knowingly misinterpreting or reusing unrelated data to trick readers into accepting an argument or claim.

These definitions are all quite specific to the *use* of data and the scientific focus on the generation and review of publications and grants and therefore must be reconsidered when non-research analyses or interpretations/inferences are the focus of work. By contrast, when considering data itself, including its access, use, and management, the legal issues have focused on protecting the privacy of the individuals from whom data are obtained (sometimes without the individuals’ knowledge or permission) and the ownership of this data (however, see, e.g., Dwork and Mulligan 2013; Steinmann et al. 2016). Since individuals are typically unable to “benefit” in any way from their own data alone, those who collect data from large numbers of individuals (e.g., through health-care systems, through purchasing, browsing, or through national data collection efforts) both expend the resources to collect the data and those required to analyze them in the aggregate. This is the basis of some claims that, although the individual may contribute their own data to a “big” data set, the owner of that data is the person or agency that collects and houses/manages that data for analysis and use.

Additional legal issues may arise when conclusions or inferences are made based on aggregated data that result in biased decisions, policies, and/or resource allocation against any group (Dwork and Mulligan 2013). Moreover, those who collect, house, and manage data from individuals incur the legal (and ethical) obligation to maintain the security and the integrity of that data – to protect the source of the data (the individuals about whom it is ultimately descriptive) and to ensure that decisions and inferences based on that data are unbiased and do not adversely affect these individuals. Claims based on big data include formally derived risk estimates, e.g., from health systems data about risk factors that can be modified or otherwise targeted to improve health outcomes in the aggregate (i.e., not specifically for any single individual) and from epidemic or pandemic trends such as influenza, Zika, and Ebola. However, they also include informal risk descriptions, such as claims made in support of the Brexit vote (2016) or climate change denial (2010–2017). False claims based on formal analyses may arise

from faulty assumptions rather than the intent to defraud or mislead and so may be difficult to label as “illegal.” Fraud relating to big data may become a legal issue in business and industry contexts where shareholders or investors are misled intentionally by inappropriate analyses; national and international commercial speech representing false claims – whether arising from very large or typical size data sets – are already subject to laws protecting consumers. Governments or government agents falsifying big data, or committing fraud, may be subject to sanctions by external bodies (e.g., see consumer price index estimation/falsification in Argentina, 2009; gross domestic product estimation in Greece, 2016) that can lead foreign investors to distrust their data or analyses. These recent examples used extant law (i.e., not new or data-specific regulations) to improperly prosecute competent analysts whose results did not match the governments’ self-image. Thus, much of the current law (nationally and internationally) relating to big data can be extrapolated for new cases, although future legal protections may be needed as more data become more widely available for the confirmation of results that some government bodies wish to conceal.

Ethical Implications/Issues

By contrast to the legal issues, ethical issues relating to big data research and practice are much less straightforward. Two major challenges to understanding, or even anticipating, ethical implications of/issues arising from the collection, use, or interpretation of big data are (1) most training on ethics for those who will eventually work in this domain are focused on *research and not practice* that will not result in peer review/publication; and (2) this training is typically considered to be “discipline specific” – based on norms for specific scientific domains.

People who work with, but do not *conduct research* in or with, big data may feel or be told – falsely – that there are no ethical issues with which they should be concerned. Because much of the training in “ethical conduct in research” relates to interactions with research

subjects and other researchers, publications, and professional conducts in research or laboratory settings, it can appear – erroneously – that these are the only ethical implications of interacting or working with big data (or *any* data).

The fact that training in ethical research practices is typically considered to be discipline specific is another worrisome challenge. Individuals from many different backgrounds are engaging in both research and non-research work with big data, suggesting that no single discipline can assert its norms or code of conduct as “the best” approach to training big data workers in ethical professional behavior. Two extremely damaging attitudes impeding addressing this growing challenge are the attitudes that (a) “ethical behavior is just common sense” and (b) whatever is not illegal is actually acceptable. In 2016 alone, two edited volumes outlined ethical considerations around research and practice with big data. Both the American Statistical Association (<http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx>) and the Association for Computing Machinery (<https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct>) have codes of ethical practice which accommodate the analysis, management, collection, and interpretation of data (big, big, or “small”), and the Markkula Center for Applied Ethics at Santa Clara University maintains (as of January 2017) a listing of ethical uses or considerations of big data <https://www.scu.edu/ethics/focus-areas/internet-ethics/articles/articles-about-ethics-and-big-data/>. A casual reading of any of these articles underscores that there is *very little* to be gleaned from “common sense” relating to ethical behavior when it comes to big data, and the major concern of groups from trade and professional associations to governments around the world is that technology may change so quickly that anything which is rendered *illegal* today may become obsolete tomorrow. However, it is widely argued (and some might argue, just as widely ignored) that what is “illegal” is only a subset of what is “unethical” in every context except the ones where these are explicitly linked (e.g., medicine, nursing, architecture, engineering). Ethical implications arise in both research and nonscience

analysis involving big data, and we cannot know or even anticipate what they can be in the future. These facts do not remove the ethical challenges or prevent their emergence. Thus, all those who are being trained to work with (big) data should receive comprehensive training in ethical reasoning to promote the identification and appropriate response to ethical implications and issues arising from work in the domain.

Conclusion

Ethical and legal implications of the collection, management, analysis, and interpretation of big data exist and evolve as rapidly as the technology and methodologies themselves. Because it is unwieldy – and essentially not possible – to consider training all big data workers and researchers in the ELSI (or just ethical, or just legal, implications) of the domain, training in reasoning and practicing with big data ethically needs to be comprehensive and integrated throughout the preparation to engage in this work. Modern work – with big data – involves multiple disciplines and is not uniquely research oriented. The norms for professionalism, integrity, and transparency arising from two key professions aligned with big data – statistics and computing – are concrete, current, consensus-based codes of conduct, and their transmission to all who will engage with big data, whether as researchers or workers, is essential.

Further Reading

- Collmann, J., & Matei, S. A. (Eds.). (2016). *Ethical reasoning in big data: An exploratory analysis*. Cham, CH: Springer International Publishing.
- Dwork, C., & Muliigan, D. K. (2013). It's not privacy, and it's not fair. *Stanford Law Review Online*, 66(35), 35–40.
- Mittelstadt, B. D., & Luciano, F. (Eds.). (2016). *The ethics of biomedical big data*. Cham, CH: Springer International Publishing.
- Steinmann, M., Shuster, J., Collmann, J., Matei, S., Tractenberg, R. E., FitzGerald, K., Morgan, G., & Richardson, D. (2016). Embedding privacy and ethical values in big data technology. In S. A. Matei, M. Russell, & E. Bertino (Eds.), *Transparency on social media – tools, methods and algorithms for mediating online interactions* (pp. 277–301). New York, NY: Springer.

Ethics

Erik W. Kuiler

George Mason University, Arlington, VA, USA

Big data ethics focus on the conduct of individuals and organizations, both public and private, engaged in the application of information technologies (IT) to the construction, acquisition, manipulation, analytics, dissemination, and management of very large datasets. In application, the purpose of big data codes of ethics is to delineate the moral dimensions of the systematic computational analyses of structured and unstructured data and their attendant outcomes and to guide the conduct of actors engaged in those activities.

The expanding domains of big data collection and analytics have introduced the potential for pervasive algorithm-driven power asymmetries that facilitate corruption and the commodification of human rights within and across such spheres as health care, education, and access to the workplace. By prejudging human beings, algorithm-based predictive big data-based analytics and practices may be used to perpetuate or increase de jure and de facto inequalities regarding access to opportunities for well-being based on, for example, gender, ethnicity, race, country of origin, caste, language, and also political ideology or religion. Related asymmetries are understood in terms of ethical obligations versus violations and are framed in terms of what should or should not occur, such that they should be eliminated. To that end, big data ethics typically are discussed along broadly practical dimensions related to methodological integrity, bias mitigation, and security and data privacy.

Method Integrity

To ensure heuristic integrity, big data analytics must meet specific ethical obligations and provide the appropriate documentation. Disclosure of research methods requires specific kinds of information, such as a *statement of the problem*, a clear

definition of a research problem, and the research goals and objectives. A *data collection and management plan* should provide a statement of the data analyzed, or to be analyzed, and methods of collection, storage, and safekeeping. Data privacy and security assurance enforcement oversight and processes should be explicitly stated, with a description of the mechanisms and processes for ensuring data privacy and security. For example, in the USA, the Health Insurance Portability and Accountability Act (HIPAA) information and Personally Identifiable Information (PII) require special considerations to ensure that data privacy and security requirements are met. In addition, a *statement of data currency* must specify when the data were, or are to be, collected. Where appropriate and applicable, *hypothesis formulation* should be clearly explained, describing the hypotheses used, or to be used, in the analysis and their formulation. Similarly, *hypotheses testing* should be explained, describing hypothesis testing paradigms, including, for example, algorithms, units of analysis, units of measure, etc. applied, or to be applied, in the research. Likewise, *results dissemination* should be explained in terms of dissemination mechanisms and processes. A statement of *reproducibility* should also be included, indicating how methods and data can be acquired and how the research can be duplicated by other analysts.

Bias Mitigation

Big datasets, by their very size, make it difficult to identify and mitigate different biases. For example, *algorithmic bias* includes systematic, repeatable errors introduced (intentionally or unintentionally) by formulae and paradigms that produce predisposed outcomes that arbitrarily assign greater value or privileges to some groups over others. *Sampling bias* refers to systematic, repeatable errors introduced by data that reflect historical inequalities or asymmetries. *Cultural bias* can be based on systematic, repeatable errors introduced by analytical paradigms that reflect personal or community mores and values, whereas *ideological bias* is systematic, repeatable errors introduced by analytical designs and

algorithms that reflect specific perspectives or dogmas. *Epistemic bias* reflects stove-piping: systematic, repeatable errors introduced by the adherence to professional or academic points of view shared within specific disciplines without exploring other perspectives external to those disciplines and the perpetuation of intellectual silos.

Data Privacy and Security Assurance

Big data, especially in cloud-based environments, require special care to assure that data security and privacy regimens are specified and maintained. Data privacy processes ensure that sensitive personal or organizational data are not acquired, manipulated, disseminated, or stored without the consent of the subjects, providers, or owners. Data security processes protect data from unauthorized access, and include data encryption, tokenization, hashing, and key management, among others.

Summary

Big data ethics guide the professional conduct of individuals engaged in the acquisition, manipulation, analytics, dissemination, and management of very large datasets. The proper application of big data ethics ensures method integrity and transparency, bias mitigation, data privacy assurance, and data security.

Further Reading

- American Statistical Association. *American Statistical Association Ethical Guidelines for Statistical Practice*. Available from: <https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx>
- Association for Computing Machinery. *ACM code of ethics and professional conduct*. Available from: <https://www.acm.org/code-of-ethics>
- Data Science Association. *Data Science Association Code of Professional Conduct*. Available from: <https://www.datascienceassn.org/code-of-conduct.html>
- Institute of Electrical and Electronics Engineers. *IEEE ethics and member conduct*. Available from: <https://www.ieee.org/about/corporate/governance/p7-8.html>

- Richterich, A. (2018). *The big data agenda: Data ethics and critical data studies*. London: University of Westminster Press.
- United States Senate. *The Data Accountability and Transparency Act of 2020 Draft*. Available from https://www.banking.senate.gov/download/brown_2020-data-discussion-draft
- Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(2), 1–6.

Ethnographic Observation

► Contexts

European Commission

Chiara Valentini

Department of Management, Aarhus University,
School of Business and Social Sciences, Aarhus,
Denmark

Introduction

The phenomenon of big data and how organizations collect and handle personal information are often discussed in relation to human rights and data protection. Recent developments in legislation about human rights, privacy matters, and data protection are taking place more and more at the European Union level. The institution that proposes and drafts legislation is the European Commission. The European Commission is one of three EU institutions in charge of policy making. It has proposal and executive functions and represents the interests of the citizens of the European Union. Specifically, it is in charge of setting objectives and political priorities for action. It proposes legislation that is approved by the European Parliament and the Council of the European Union. It oversees the management and implementation of EU policies and the EU budget. Together with the European Court of Justice, it enforces European law, and it represents the EU outside the European

Union zone, for example, in negotiating trade agreements between the EU and other countries (Nugget 2010). It has its headquarters in Brussels, Belgium, but has offices also in Luxembourg. The European Commission is also present with own representative offices in each EU member state. The representations of the European Commission can be considered the “eyes” and “ears” of the Commission at the local level providing the headquarters with updated information on major issues of importance occurring in each member state.

Election Procedure and Organizational Structure

The European Commission is formally a college of commissioners. Today, it comprises 28 commissioners, including the President and the Vice-Presidents (European Commission, 2017a). The commissioners are in charge of one or more portfolios, that is, they are responsible for specific policy areas (Nugget 2010).

Until 1993, the European Commission was appointed every 4 years by common accord of governments of member states, and initially the number of commissioners reflected the number of states in the European Community. After the introduction of the Treaty of Maastricht in 1993, the length of the mandate and election procedures was revised. The European Commission mandate was changed to 5 years with the college of commissioners appointed 6 months after the European Parliament elections. Furthermore, the composition of the European Commission has to be negotiated with the European Parliament. The candidate of the President of the European Commission is also chosen by the governments of EU member states in consultation with the European Parliament. The commissioners, which were in the past nominated by the governments of member states, are chosen by the President of the European Commission. Once the college of commissioners is formed, it needs to get its approval from the Council of the European Union and the European Parliament (Nugget 2010). The position of the European Parliament in influencing the

composition of the college of commissioners and the election of the president of the college of commissioners, that is, the President of the European Commission, was further strengthened with subsequent treaties. The latest treaty, the Lisbon Treaty, also stipulated that one of the commissioners should be the person holding the post of High Representative of the Union for Foreign Affairs and Security Policy. This position somehow resembles that of a Minister of Foreign Affairs, yet, with more limited powers.

The main representative of the European Commission with other EU institutions and with external institutions is the President. While all decisions made in the European Commission are collective, the President's main role is to give a sense of direction to the commissioners. He or she allocates commissioners' portfolios, has the power to lay off commissioners from their post, and is directly responsible for the management of the Secretariat General which is in charge of all activities in the Commission. The President also maintains relations with the other two decision-making institutions, that is, the European Parliament and the Council of the European Union, and can assume specific policy responsibilities of his/her own initiative (Nugget 2010).

The college of commissioners represents the interests of the entire union. Commissioners are, thus, asked to be impartial and independent from the interests of their country of origin in performing their duties. Commissioners are generally national politicians of high rank, often former national ministers. They hold one or more portfolios. Prior to the implementation of the Amsterdam Treaty, when the President of the European Commission gained more power to decide which commissioners should hold which portfolio, the distribution of portfolios among commissioners was largely a matter of negotiation between national governments and of political balance among the member states (Nugget 2010).

Each commissioner has his/her own cabinet that helps to perform different duties. While originally civil servants working in a commissioner's cabinet came from the same country of the commissioner, from the late 1990s, each cabinet was

required to have civil servants of at least three different nationalities. This decision was made to prevent that specific national interests dominate the discussion on policy developments. The cabinets perform research and policy analyses that are essential in keeping the commissioners informed on developments of the assigned policy areas but also helping the commissioners to be updated on other cabinets and commissioners' activities.

The European Commission's Legislative Work

Administratively speaking, the European Commission is divided into Directorate-Generals (DGs) and other services which are organizational units specializing in specific policy areas and services. According to the European Commission, over 32,500 civil servants work for the European Commission in one of these units in summer 2017 (European Commission 2017b). The European Commission's proposals are prepared by one of these DGs. Drafts of proposals are crafted by middle-ranking civil servants in each DGs. These officers often rely on outside assistance, for instance, from consultants, academics, national experts, officials, and interest groups, too. Draft proposals are scrutinized by the Secretariat General to meet the existing legal requirements and procedures. The approved draft is then inspected by senior civil servants in the DGs, the cabinet personnel, and finally reaches the commissioners. The draft proposal is shaped and revised continuously during this process. Once the college of commissioners meets to discuss and approve the draft, they may accept it in the submitted form, reject it, or ask for revisions. If revisions are asked, the draft goes back to the responsible DG (Nugget 2010).

The European Commission's proposals become official only once the college of commissioners adopts them. The decisions are taken by consensus but majority voting is possible. Typically, the leadership for making proposals pertaining specific policy areas lies on the commissioner holding the portfolio in question.

Proposals related to fundamental rights, data protection, and citizens' justice are typically carried out by the DG for Justice. Because political, social, and economic issues may affect several policy areas, informal and ad hoc consultations between different commissioners who may be particularly affected by a proposal occur. There are also groups of commissioners in related and overlapped policy areas that facilitate cooperation across the DGs and enable the discussion when the college of commissioners meets.

The European Commission's Position Toward Big Data

The European Commission released a position document in summer 2014 in response to the European Council's call for action in autumn 2013 to develop a single market for big data and cloud computing. The European Commission is overall positive toward big data and sees data at the center of the future knowledge economy and society. At the same time, it stresses that an unregulated use of data can undermine fundamental rights. To develop a digital economy in Europe as recommended by the European Council, the Commission has proposed a framework to boost data-driven innovations through the creation of infrastructures that allow for quality, reliable, and interoperable datasets. Among others, the Commission seeks to improve the framework conditions that facilitate value generation from datasets, for example, by supporting the creation of collaborations among different players such as universities/public research institutes, private entities, and businesses (European Commission 2014b). The European Commission actions to achieve these goals revolve around five main initiatives. First, it intends to create a European public-private partnership on data and to develop digital entrepreneurship. Second, the Commission seeks to develop an open data incubator, that is, programs designed to support the successful development of entrepreneurial companies. Third, it aims at increasing the development of specific competences necessary to have more skilled data professionals who are considered important for

developing a knowledge economy. Fourth, it plans to establish data market monitoring tools to measure the size and trends of the European data market. Finally, it intends to consult and engage stakeholders and research communities across industries and fields to identify major priorities for research and innovation in relation to digital economy (European Commission 2014b).

Since 2014, the European Commission has worked in developing frameworks that promote open data policy, open standards, and data interoperability. To secure that the development of big data initiatives does not undermine fundamental rights to personal data protection, the European Commission has worked on revising EU legislation. A revised version of the 2012 proposal for data protection regulation was approved by the European Parliament and by the Council of the European Union in 2014. To guarantee security and data protection as well as to support organizations in implementing big data initiatives, the Commission intends to work with each member state and relevant stakeholders to inform and guide private organizations on issues related to data collection and processing such as data anonymization and pseudonymization, data minimization, and personal data risk analysis. The Commission is also interested in enhancing consumer awareness on big data and their data protection rights (European Commission 2014b).

In relation to cloud computing, the Commission initiated in 2012 a strategy to achieve a common agreement on standardization guidelines for cloud computing service. Relevant stakeholders and industry leaders were invited to discuss and propose guidelines in 2013. A common agreement was reached and guidelines have been published in June 2014. Despite the diversity of technologies, businesses, and national and local policies, the guidelines aim at facilitating the comparability of service level agreements in cloud computing, providing clarity for cloud service customers, and generating trust in cloud computing services (Watts et al. 2014, p. 596) and are thus considered a step forward in regulating data mining and processing in Europe but also around the world.

The European Commission's Concerns on Big Data

The European Commission has recognized that the big data phenomenon offers several opportunities for growth and innovation, yet there are no sufficient measures to safeguard data privacy and security of users and consumers. A number of challenges have been identified including the anonymization of data, that is, data which is deprived of direct or indirect personal identifiers, and the problem of creating personal profiles based on patterns of individuals' behaviors obtained through data mining algorithms. The creation of such personal profiles could be misused and affect how individuals are treated. The European Commission is particularly concerned with the aspect of maintaining the fundamental rights of citizens and of possible misuse of EU citizens' personal information by data collectors and through data processing. There are also a number of security risks, particularly when data collected by a device are transferred elsewhere.

The lack of transparency on when and how data is collected and for which purposes are the major concerns. Countless data from European citizens is collected and processed by non-EU companies and/or outside the EU zone. Therefore, various types of data are handled in many different manners, and existing EU legislation is short in regulating possible abuses and issues occurring beyond its jurisdiction. The European Commission has already initiated negotiations with established trade partners such as the United States. Since March 2011, the European Commission has been negotiating procedures for transferring personal data from the EU to the US for law enforcement purposes. The European Commission does not consider the *US-EU Safe Harbor Framework* a sufficient measure to protect fundamental rights such as data privacy and security of Europeans (European Commission 2012). The *US-EU Safe Harbor Framework*, approved in 2000, is a framework that allows registered US organizations to have access to data from EU citizens upon declaring that they have adequate privacy protection mechanisms in place (Export.gov 2013). The concern of data security and privacy has increased after the revelations of the

National Security Agency (NSA) data collection practices on Europeans and subsequent statements by the NSA that data on European citizens was supplied by European intelligence services according to existing agreements (Brown 2015). The European Commission has initiated a process for regulating data processing. With the United States, an agreement called *data protection umbrella agreement* has been reached in certain areas. The agreement deals with personal data such as names, addresses, and criminal records transferred from the EU to the US for reasons of prevention, detection, investigation, and prosecution of criminal offences, including terrorism. Yet, the EU and US had for quite some time different opinions on the right of effective judicial redress that should be granted by the US to EU citizens not resident in the United States (European Commission 2014a).

On 2 June 2016, the EU and the US agreed on a wide-ranging high-level data protection framework for criminal law enforcement cooperation, called "umbrella agreement." The agreement should improve, in particular, EU citizens' rights by providing equal treatment with US citizens when it comes to judicial redress rights before US courts (CEU 2016).

Critics further point out that the European Commission regulatory proposal for data protection lacks the capacity to safeguard fully European citizens' rights. Specifically, the problem of anonymization of data is considered by some not well addressed by the European Commission's proposal, since data that is stripped away of names and direct identifiers can still be associated to specific individuals using a limited amount of publicly available additional data (Aldhouse 2014). This could leave space to commercial entities and third-party vendors for exploiting the potential of big data while legally abiding to the EU legislation.

Cross-References

- [Big Data Theory](#)
- [Cloud Computing](#)
- [Cloud Services](#)
- [Data Mining](#)

- [Data Profiling](#)
- [European Commission: Directorate-General for Justice \(Data Protection Division\)](#)
- [European Union](#)
- [National Security Administration \(NSA\)](#)
- [Open Data](#)
- [Privacy](#)

Further Reading

- Aldhouse, F. (2014). Anonymisation of personal data – A missed opportunity for the European Commission. *Computer Law and Security Review*, 30(4), 403–418.
- Brown, I. (2015). The feasibility of transatlantic privacy-protective standards for surveillance. *International Journal of Law and Information Technology*, 23(1), 23–40.
- CEU. (2016, June 2). Enhanced data protection rights for EU citizens in law enforcement cooperation: EU and US sign “Umbrella agreement”. *Press Release of the Council of the European Union*. <http://www.consilium.europa.eu/en/press/press-releases/2016/06/02-umbrella-agreement/>. Accessed 7 July 2016.
- European Commission. (2012, October 9). How will the ‘safe harbor’ arrangement for personal data transfers to the US work? http://ec.europa.eu/justice/policies/privacy/thridcountries/adequacy-faq1_en.htm. Accessed 21 Oct 2014.
- European Commission. (2014a, June). *Factsheet EU-USA negotiations on data protection*. http://ec.europa.eu/justice/data-protection/files/factsheets/umbrella_factsheet_en.pdf. Accessed 21 Oct 2014.
- European Commission. (2014b, July 2). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions. Towards a thriving data-driven economy. *COM(2014) 442 final*, http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=6210. Accessed 21 Oct 2014.
- European Commission. (2017a). *The Commissioners. The European Commission's political leadership*, https://ec.europa.eu/commission/commissioners/2014-2019_en. Accessed 5 September 2017.
- European Commission. (2017b). *Staff members*. http://ec.europa.eu/civil_service/docs/hr_key_figures_en.pdf. Accessed 05 September 2017.
- Export.gov. (2013, December 18). *U.S.-EU safe harbor overview*. <http://export.gov/safeharbor/eu/>. Accessed 19 Oct 2014.
- Nugget, N. (2010). *The government and politics of the European Union* (7th ed.). New York: Palgrave Macmillan.
- Watts, M., Ohta, T., Collis, P., Tohala, A., Willis, S., Brooks, F., Zafar, O., Cross, N., & Bon, E. (2014). EU update. *Computer Law and Security Review*, 30(5), 593–598.

European Commission: Directorate-General for Justice (Data Protection Division)

Chiara Valentini

Department of Management, Aarhus University,
School of Business and Social Sciences, Aarhus,
Denmark

Introduction

Debates on big data, data-related situations and international developments have increased in recent years. At the European level, the European Commission has created a subunit within its Directorate-General for Justice to study and monitor the big data phenomenon and its possible impact on legislative matters. The Directorate-General for Justice is one of the main departments of the European Commission, specializing in promoting justice and citizenship policies and protecting fundamental rights. The Directorate is in charge of the justice portfolio and proposes legislative documents in relation to four areas: civil justice, criminal justice, fundamental rights and union citizenship, and equality. The justice portfolio is relatively new; it was only created in 2010 under the leadership of the President José Manuel Barroso. Previously, it was part of the former Directorate-General for Justice, Freedom and Security which was split into two departments, the Directorate-General Home Affairs and the Directorate-General for Justice (European Commission 2014).

The Data Protection Division

The Data Protection Division is a subunit of the Directorate-General for Justice specializing in all aspects concerning the protection of individual data. It provides the Directorate-General for Justice and the European Commission with independent advice on data protection matters and helps with the development of harmonized policies on

data protection in the EU countries (CEC 1995). Data protection has become an important new policy area for the Directorate-General for Justice following the approval and implementation of the 1995 EU Data Protection Directive.

The 1995 directive that came into force in 1998 has been considered the first most successful instrument for protecting personal data (Bennett and Raab 2006; Birnhack 2008). The directive binds EU member states and three members of the European Economic Area (Iceland, Lichtenstein, Norway) to establish mechanisms to monitor how personal data flows across countries within the EU zone but also in and out of third countries. It requires authorities and organizations collecting data to have in place adequate protection mechanisms to prevent misuse of sensitive data (Birnhack 2008). The directive restricts the capacity of organizations to collect any type of data. The processing of special categories of data on racial background, political beliefs, health conditions, or sexual orientation is, for example, prohibited. The directive has also increased the overall transparency of the data collection procedures, by expanding people's rights to know who gathers their personal information and when (De Hert and Papakonstantinou 2012). Authorities and organizations that intend to collect personal data have to notify individuals of their collection procedures and their data use. Individuals have the right to access the data collected and can deny certain processing. They have also the right not to be subjected to an automated decision, that is a decision that is relegated to computers which gather and process data as well as suggest or make decisions silently and with little supervision (CEC 1995).

Due to rapid technological developments and the increased globalization of many activities, the European Commission started a process of modernization of the principles constituting the 1995 directive. First, in 2001 the regulation 45/2001 on data processing and its free movement in the EU institutions was introduced. This regulation aims at protecting individuals' personal data when the processing takes place in the EU institutions and bodies. Then in 2002 a new directive on privacy and electronic communications was set to ensure

the protection of privacy in the electronic communications sector. The 2002 directive was amended in 2006 to include also aspects related to the retention of data generated or processed in connection with the provision of publicly available electronic communications services, which are services provided by means of electronic signals over, for example, telecommunications or broadcasting networks, or of public communication networks (CEC 2006). In 2008 the protection was extended to include data collection and sharing within police and judicial cooperations in criminal matters.

Latest Developments in the Data Protection Regulation

Due to the increased development, use, and access of the Internet and other digital technologies by individuals, companies, and authorities around the world, new concerns about privacy rights have called the attention of the European Commission. The 1995 directive and the following directives were considered not sufficient to provide the legal framework for protecting Europeans' fundamental rights (Birnhack 2008; De Hert and Papakonstantinou 2012). Furthermore, the 1995 directive allowed member states a certain level of freedom in the methods and instruments implementing EU legislation. As a result, the Data Protection Directive was often transposed in national legislations in very different manners arising enforcement divergences (De Hert and Papakonstantinou 2012). The fragmentation of data protection legislation and the administrative burden of handling all member states' different rules motivated the EU Commissioner responsible for the Directorate-General for Justice to propose a unified, EU-wide solution in the form of a regulation in 2012 (European Commission 2012).

The 2012 proposal includes a draft EU legislative framework comprising a regulation on general data protection directly applicable to all member states and a directive specifically for personal data protection that leaves discretions to member states to decide the form and method of application. It also proposes the establishment of

specific bodies that oversee the implementation and respect of data protection rules. These are a data protection officer (DPO) to be located in every EU institution and a European data protection supervisor (EDPS). The DPO is in charge of monitoring the application of the regulation within the EU institutions whereas the EDPS has the duty of controlling the implementation of data protection rules across member states (European Commission 2013).

The initial draft was revised several times to meet the demands of the European Parliament and the Council of the European Union, the two EU decision-making institutions. In spring 2014, the European Parliament supported and pushed for a voting on the Data Protection Regulation, which is an updated version of the regulation that was first proposed by the European Commission in 2012. The final approval required, however, the support of the other two institutions.

On 15 December 2015, the European Parliament, the Council, and the Commission reached an agreement on the new data protection rules, and on 24 May 2016, the regulation entered into force, but its application is not expected before 25 May 2018 to allow each member state to transpose it into their national legislations (European Commission 2016). The approved version extends the territorial scope of its application, which means that the regulation will apply to all data processing activities concerning EU citizens even if the data processing does not take place in the European Union. The regulation includes an additional provision concerning processing children personal data and moves the responsibility to the data controller, that is the organization or authority collecting it, to prove that consent of gathering and handling personal data was given. Another amendment that the European Parliament has requested and obtained is the introduction of a new article about transferring disclosures that are not authorized by the European Union Law. This article allows organizations and authorities that collect personal data to deny releasing information to non-European law enforcement bodies for reasons that are considered to be contrary to

the European data protection principles. The “right to be forgotten,” which is the right by individuals to see removed irrelevant or excessive personal information from search engine results, is also included (Rees and Heywood 2014).

Possible Impact of the Data Protection Regulation

Critics noted that the introduction of the data protection regulation may affect substantially third-party vendors and those organizations that use third-party data, for example, for online marketing and advertising purposes. The regulation will demand data collectors who track users on the web, their pages visited, the amount of time spent in each page, and any other online movement, to prove that they have obtained individuals’ consents to use and sell personal data, otherwise they will have to pay high infringements fines. This regulation may impact the activities of multinational companies and international authorities, since it is expected that the new EU data protection standards apply to any data collected on EU citizens, no matter where data is processed. Google, Facebook, and other Internet companies have lobbied against the introduction of this data protection regulation but with little success (Chen 2014). The EU debate on data protection regulation seems to have sparked international debates on data protection in other non-EU countries and on the fitness of their national regulation. For instance, not long after the European Parliament voted for the General Data Protection Regulation, the state of California, in the U.S., passed a state law that requires technological companies to remove material posted by a minor, if the user requests it (Chen 2014).

Cross-References

- [Big Data Theory](#)
- [Charter of Fundamental Rights \(EU\)](#)
- [European Commission](#)

- European Union
- Privacy

Further Reading

- Bennett, C. J., & Raab, C. D. (2006). *The governance of privacy: Policy instruments in global perspective*. Cambridge, MA: MIT Press.
- Birnhack, M. D. (2008). The EU data protection directive: An engine of a global regime. *Computer Law and Security Review*, 24(6), 508–520.
- CEC. (1995, November 23). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Union*, L281. <http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046>.
- CEC. (2006, April 13). Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC. *Official Journal of the European Union*, L105/54. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:en:HTML>.
- Chen, F. Y. (2014, May 13). European court says Google must respect ‘right to be forgotten’. *Reuters US Edition*. <http://www.reuters.com/article/2014/05/13/us-eu-google-dataprotection-idUSBREA4C07120140513>. Accessed 8 Oct 2014.
- De Hert, P., & Papakonstantinou, V. (2012). The proposed data protection regulation replacing directive 95/46/EC: A sound system for the protection of individuals. *Computer Law and Security Review*, 28(2), 130–142.
- European Commission. (2012, January 25). Commission proposes a comprehensive reform of data protection rules to increase users’ control of their data and to cut costs for businesses. Press Release. http://europa.eu/rapid/press-release_IP-12-46_en.htm. Accessed 10 Oct 2014.
- European Commission. (2013, July, 16). *European data protection supervisor*. http://ec.europa.eu/justice/data-protection/bodies/supervisor/index_en.htm. Accessed 10 Oct 2014.
- European Commission. (2014). *Policies and activities. DG Justice*. http://ec.europa.eu/justice/index_en.htm#newsroom-tab. Accessed 10 Oct 2014.
- European Commission. (2016, July 6). *Reform of EU data protection rules*. http://ec.europa.eu/justice/data-protection/reform/index_en.htm. Accessed 7 July 2016.
- Rees, C., & Heywood, D. (2014). The ‘right to be forgotten’ or the ‘principle that has been remembered’. *Computer Law and Security Review*, 30(5), 574–578.

European Union

Chiara Valentini

Department of Management, Aarhus University,
School of Business and Social Sciences, Aarhus,
Denmark

Introduction

The development and integration of big data concerns legislators and governments around the world. In Europe, legislation regulating big data and initiatives promoting the development of digital economy are handled at the European Union level. The European Union (EU) is a union of European member states. It was formally established in 1993 when the Maastricht Treaty came into force to overcome the limits of the European Community and strengthen the economic and political agreements of participating countries. The European Community was established in 1957 with the Treaty of Rome and had primarily an economic purpose to establish a common market among six nation-states—Belgium, France, West Germany, Italy, Luxembourg, and the Netherlands. The EU is a supranational polity that acts in some policy areas as a federation, that is, its power is above member states’ legislation, and in other policy areas as a confederation of independent states, similar to an inter-governmental organization, that is, it can provide some guidelines but decisions and agreements are not enforceable, and member states are free to decide whether or to which extent to follow them (Valentini 2008). Its political status is thus unique in several respects, because nation-states that join the European Union must accept to relinquish part of their national power in return for representation in the EU institutions. The EU comprises diverse supranational independent institutions such as the European Commission, the European Parliament, and the Council of the European Union also known as the Council of Ministers. It also operates through

intergovernmental negotiated decisions by member states that gather together, for instance, in the European Council.

The supranational polity has grown from the six founding European countries to current 27, after the United Kingdom decided to leave the EU in summer 2016. On January 2016, the population of the EU was about 510 million people (Eurostat 2016). To become an EU member state, countries need to meet the so-called “Copenhagen criteria”. These require that a candidate country has achieved institutional stability guaranteeing democracy, is based on the Rule of Law, has in place policies for protecting human and minority rights, has a functioning market economy, and can cope with competitive pressures and market forces (Nugget 2010, p. 43). Five countries are recognized as candidates for membership: Albania, Macedonia, Montenegro, Serbia, and Turkey. Other countries, such as Iceland, Lichtenstein, Norway, and Switzerland, are not EU members but are part of the European Free Trade Association and thus enjoy specific trade agreements (EFTA 2014).

EU Main Institutions

The EU political direction is set by the European Council which has no legislative powers but acts as body and issues guidelines to the European Commission, the European Parliament, and the Council of the European Union. It comprises a President, the national heads of state or government and the President of European Commission. The three main institutions involved in the legislative process are the European Commission, the European Parliament, and the Council of the European Union. The European Commission is the institution that drafts and proposes legislation based on its own initiative but also on suggestions made by the European Council, the European Parliament, the Council of the European Union, or other external political actors. It comprises a President and 27 commissioners who are responsible each of one or more policy areas. The Commission is also responsible for monitoring the

implementation of the EU legislation once adopted (Nugget 2010).

The European Parliament is elected every five years with direct universal suffrage since 1979. It is the only EU institution that is directly elected by citizens aged 18 years or older in all member states, except Austria where the voting age is 16. Voting is compulsory in four member states (Belgium, Luxembourg, Cyprus, and Greece), and European citizens that reside in a member state other than their own have the right to vote for the European Parliament elections in their state of residence (European Parliament 2014). The European Parliament comprises a President and 751 members across seven political groups representing the left, central, and right political positions. In the co-decision procedure, that is, the most common procedure for passing EU law, the Parliament together with the Council of the European Union is in charge of approving EU legislation.

The Council of the European Union represents the executive governments of the EU’s member states and comprises a Presidency and a council of 27 ministers (one per member state) that changes according to the policy area under discussion. There are ten different configurations, that is, variations of council composition. The Presidency is a position held by a national government and rotates every 6 months among the governments of the member states. To maintain some consistency in the program, the Council has adopted an agreement called *Presidency trios* under which three successive presidencies share common political programs. The administrative activities of the Council of the European Union are run by the Council’s General Secretariat.

Decision-making in the Council can be by unanimity, by qualified majority (votes are weighted by the demographic clause, which means that high-populated countries have more votes than low-populated ones), and by simple majority (Nugget 2010).

The legal power is given to the Court of Justice which makes sure that the EU law is correctly interpreted and implemented in member states. The Court of Auditors checks cases of maladministration in the EU institutions and bodies.

Typically maladministration cases that are solicited by citizens, business, and organizations are handled by the European Ombudsman. Citizens' privacy issues are handled, on the other hand, by the European Data Protection Supervisor who is in charge of safeguarding the privacy of people's personal data.

While all EU member states are part of the single market, only 19 of them have joined the monetary union by introducing a single currency, the euro. The EU institution responsible for the European monetary policy is the European Central Bank (ECB). Located in Frankfurt, Germany, the ECB's main role is to maintain price stability. It also defines and implements the monetary policy, conducts foreign exchange operations, holds and manages the official foreign reserves of the euro area countries, and promotes smooth operations of payment systems (Treaty of the European Union 1992, p. 69). The Union has also its own specific international financial institution, the European Investment Bank (EIB), publicly owned by all EU member states. The EIB finances EU investment projects and helps small businesses through the European Investment Fund (EIB 2013).

Another important EU institution is the European External Action Service which is a unit supporting the activities of the High Representative of the Union for Foreign Affairs and Security Policy, and its main role is to ensure that all diplomacy, trade, development aid, and work with global organization activities that the EU undertakes are consistent and effective (EEAS 2014). Other interinstitutional bodies that play a role in the activities of the EU are the European Economic and Social Committee representing civil society, employers, and employees through a consultative assembly and issuing opinions to the EU institutions, and the Committee of Regions representing the interests of regional and local authorities in the member states (Nugget 2010).

Big Data and the EU

The EU position toward big data is generally positive. It believes that data can create enormous

value for the global economy, driving innovation, productivity, efficiency, and growth (Tene and Polonesky 2012). The EU particularly sees in big data opportunities to improve public services for citizens such as healthcare, transportation, and traffic regulation. It also believes that big data can increase innovation and clearly expresses an interest in further developing its use. In the last five years the EU has promoted the creation of a "cloud of public services" for the delivery of more flexible public services. These can be provided by combining building blocks, such as IT components for building e-government services involving ID, signatures, invoices, data exchange, and machine translation and by allowing service sharing between public and private providers (Taylor 2014). EU Open Data rules were approved in spring 2013, and in the next years, it is expected that the new rules will make all public sector information across the EU available for reuse, provided that the data is generally accessible and not personal (European Commission 2013). The EU believes that better regulations that protect citizens' rights as well as a better framework to help organizations taking advantage of big data are top priorities in the Europe 2020 digital agenda.

Even today with new challenges related to collecting data across border through mobile and other smart technologies, the EU and the US positions tend to traditionally differ in relation to data protection. The EU has been cooperating closely with US law enforcement agencies to share information about online behavior in order to identify terrorists and other criminals. Among the existing agreement, the EU and US have information-sharing arrangements for their police and judicial bodies, two treaties on extradition and mutual legal assistance and accords on container security and airline passenger data. Yet, the EU and US have different opinions on data privacy and data protection (Archick 2013, September 14). The 1995 Data Protection Directive, that was the main regulation in the EU until 2016, preventing the personal data of individuals living in the EU from being shared with anyone else without express consent, was perceived by the US counterpart as too restrictive and undermining

free market economy. Salbu (2002) noted that the EU 1995 Directive had negative impacts on global negotiation because companies had to comply with the EU requirements and these can be more restrictive than in other countries. Scholars observe that the 1995 Directive was not coercive since countries outside the EU were not asked to change their laws to fit the directive. Yet, as Birnhack (2008) noted, countries that wished to engage in data transactions with EU member states were indirectly required to provide an adequate level of data protection.

The 1995 Data Protection Directive was considered to be one of the strictest data protection legislations in the world. Yet, different scandals and the increased concern by citizens on how their personal data is handled by organizations (Special Eurobarometer 359 2011) have brought the issue of privacy and security on the top EU political agenda. The European Parliament commissioned an investigation on the status of the EU Intelligence and identified that several countries have in place mass surveillance programs in which data collection and processing activities go beyond monitoring specific individuals for criminal or terrorist reasons. Milaj and Bonnici (2014) argue that mass surveillance not only damages the privacy of citizens but limits the guarantees offered by the principle of presumption of innocence during the stages of a legal process. Similarly, Leese (2014) argues that pattern-based categorizations in data-driven profiling can impact the EU's non-discrimination framework, because possible cases of discrimination will be less visible and traceable, leading to diminishing accountability.

As results of these increasing concerns, in 2013 the EU launched a cybersecurity strategy to address shortcomings in the current system. The network information security (NIS) directive, adopted by the European Parliament on 6 July 2016 (European Commission 2017b, May 9), requires all member states to set up a national cybersecurity strategy including Computer Emergency Response Teams (CERTs) to react to attacks and security breaches. On September 2012, the EU decided to set up a permanent Computer Emergency Response Team (CERT-EU) for the EU institutions, agencies, and bodies comprising IT security experts from the main

EU institutions. The CERT-EU cooperates closely with other CERTs in the member states and beyond as well as with specialized IT security companies (CERT-EU 2014). Furthermore, the EU revised the 1995 Data Protection Directive and approved a new regulation, the General Data Protection Regulation (GDPR). GDPR entered into force on 24 May 2016, but it will apply from 25 May 2018 to allow each member state to transpose the directive into own national legislation (European Commission 2017a). GDPR poses a number of issues for international partners such as the US. These must abide to GDPR if personal information on EU citizens is collected by any organization or body regardless of whether it is located in the EU or not. This means that cross-border transfer of EU citizens' personal data outside of the EU is only permissible when GDPR conditions are met. In practice, the entrance into force of GDPR will require organizations collecting EU citizens' personal data to have a Data Protection Officer and to conduct privacy impact assessments to ensure they comply with the regulation in order to avoid being subject of substantial fines. GDPR is considered one of the most advanced data protection regulations in the world, yet it is still to be seen if it benefits or hampers the EU capacity of taking advantage of the opportunities that big data can offer.

Cross-References

- Cloud Computing
- Data Mining
- European Commission
- European Commission: Directorate-General for Justice (Data Protection Division)
- Metadata
- National Security Administration (NSA)
- Privacy

Further Reading

Archick, K. (2013, September 14). U.S.-EU cooperation against terrorism. *Congressional Research Service* 7-5700. <http://fas.org/sgp/crs/row/RS22030.pdf>. Accessed 31 Oct 2014.

- Birnhack, M. D. (2008). The EU data protection directive: An engine of a global regime. *Computer Law and Security Review*, 24(6), 469–570.
- CERT-EU. (2014). About us. http://cert.europa.eu/cert/plainedition/en/cert_about.html. Accessed 31 Oct 2014.
- EEAS. (2014). *The EU's many international roles. European Union External Action*. http://www.eeas.europa.eu/what_we_do/index_en.htm. Accessed 30 Oct 2014.
- EFTA. (2014). The European free trade association. <http://www.efta.int/about-efta/european-free-trade-association>. Accessed 30 Oct 2014.
- EIB. (2013). Frequently asked questions. <http://www.eib.org/infocentre/faq/index.htm#what-is-the-eib>. Accessed 30 Oct 2014.
- European Commission. (2007). *Framework for advancing transatlantic economic integration between the European Union and the United States of America*. http://trade.ec.europa.eu/doclib/docs/2007/may/tradoc_134654.pdf. Accessed 21 Oct 2014.
- European Commission. (2013). *Commission welcomes parliament adoption of new EU open data rules*. Press Release. http://europa.eu/rapid/press-release_MEMO-13-555_en.htm. Accessed 30 Oct 2014.
- European Commission. (2017a). *Protection of personal data*. <http://ec.europa.eu/justice/data-protection/>. Accessed 7 Sep 2017.
- European Commission. (2017b). *The Directive on security of network and information systems (NIS Directive). European Commission, Strategy, Single Market*. <https://ec.europa.eu/digital-single-market/en/network-and-information-security-nis-directive>. Accessed 7 Sep 2017.
- European Parliament. (2014). *The European parliament: Electoral procedures*. Factsheet on the European Union. http://www.europarl.europa.eu/ftu/pdf/en/FTU_1.3.4.pdf. Accessed 24 Oct 2014.
- Eurostat. (2016). Population on 1 January. <http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&plugin=1&language=en&pcode=tps00001>. Accessed 07 July 2016.
- Leese, M. (2014). The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue*, 45(5), 494–511.
- Milaj, J., & Bonnici, J. P. M. (2014). Unwitting subjects of surveillance and the presumption of innocence. *Computer Law and Security Review*, 30(4), 419–428.
- Nugget, N. (2010). *The government and politics of the european union* (7th edn). New York: Palgrave Macmillan.
- Salbu, S. R. (2002). The European union data privacy directive and international relations. *Vanderbilt Journal of Transnational Law*, 35, 655–695.
- Special Eurobarometer 359. (2011). Attitudes on data protection and electronic identity in the European Union. Report. *Gallup for the European Commission*. http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en.pdf. Accessed 30 Oct 2014.
- Taylor, S. (2014, June). Data: The new currency? European Voice. <http://www.europeanvoice.com/research-papers/>. Accessed 31 Oct 2014.
- Tene, O., & Polonesky, J. (2012, February 2). Privacy in the age of big data. A time for big decisions. *Stanford Law Review Online*. <http://www.stanfordlawreview.org/online/privacy-paradox/big-data>. Accessed 30 Oct 2014.
- Treaty of the European Union. (1992). *Official journal of the European Community*. https://www.ecb.europa.eu/ecb/legal/pdf/maastricht_en.pdf. Accessed 30 Oct 2014.
- Valentini, C. (2008). *Promoting the European Union. comparative analysis of EU communication strategies in Finland and in Italy*. Doctoral dissertation. Jyväskylä Studies in Humanities, 87. Finland: University of Jyväskylä Press.

European Union Data Protection Supervisor

Catherine Easton

School of Law, Lancaster University,
Bailrigg, UK

The European Union Data Protection Supervisor (EDPS) is an independent supervisory authority established by Regulation (EC) No 45/2001 on the processing of personal data. This regulation also outlines the duties and responsibilities of the authority which, at a high level, focuses upon ensuring that the institutions of the European Union uphold individuals' fundamental rights and freedoms, in particular the right to privacy. In this way the holder of the office seeks to ensure that European Union provisions regarding data protection are applied, and measures taken to achieve compliance are monitored. The EDPS also has an advisory function and provides guidance to the EU institutions and data subjects on the application of data protection measures. The European Parliament and Council, after an open process, appoint the supervisor and the assistant supervisor both for periods of 5 years. Since 2014 Giovanni Buttarelli has carried out the role with Wojciech Wiewiórowski as his assistant.

Article 46 of Regulation 45/2001 outlines in further detail the duties of this authority, in

addition to those outlined above: hearing and investigating complaints; conducting inquiries; cooperating with national supervisory bodies and EU data protection bodies; participating in the Article 29 working group; determining and justifying relevant exemptions, safeguards, and authorizations; maintaining the register of processing operations; carrying out prior checks of notified processing; and establishing his or her own rules of procedure.

In carrying out these duties, the authority has powers to, for example, order that data requests are complied with, give a warning to a controller, impose temporary or permanent bans on processing, refer matters to another EU institution, and intervene in relevant actions brought before the European Union's Court of Justice. Each year the EDPS produces a report on the authority's activities; this is submitted to the EU institutions and made available to the public.

The EDPS was consulted in the preparatory period before the EU's recent wide-ranging reform of data protection and published an opinion on potential changes. The EU's General Data Protection Regulation was passed in 2016 with the majority of its provisions coming into force within 2 years. This legislation outlines further provisions relating to the role of the EDPS; in its Article 68, it creates the European Data Protection Board, upon which the EDPS sits and also provides a secretariat.

The authority of the EDPS is vital in ensuring that the rights of citizens are upheld in this increasingly complex area in which technology is playing a fundamental role. By holding the EU institutions to account, monitoring and providing guidance, the EDPS has maintained an active presence in developing and enforcing privacy-protecting provisions across the EU.

Evidence-Based Medicine

David Brown^{1,2} and Stephen W. Brown³

¹Southern New Hampshire University, University of Central Florida College of Medicine, Huntington Beach, CA, USA

²University of Wyoming, Laramie, WY, USA

³Alliant International University, San Diego, CA, USA

Evidence-based medicine (EBM) and the term evidence-based medical practice (EBMP) are two interrelated advances in medical and health sciences that are designed to improve individual, national, and world health. They do this by conducting sophisticated research to document treatment effectiveness and by delivering the highest-quality health services. The term evidence-based medicine refers to a collection of the most up-to-date medical and other health procedures that have scientific evidence documenting their efficacy and effectiveness. Evidence-based medical practice is the practice of medicine and other health services in a way that integrates the health-care provider's expertise, the patient's values, and the best evidence-based medical information. Big data plays a central role in all aspects of both EBM and EBMP.

EBM information is generated following a series of procedure known as clinical trials. Clinical trials are research-based applications of the scientific method. The first step of this method involves the development of a research hypothesis. This hypothesis is a logically reasoned speculation that a specific medication or treatment will have a positive outcome when applied for the treatment of some specific health malady in some specific population. Hypotheses are typically developed by reviewing the literature in recent health and medical journals and by using creative thinking to generate a possible new application of the reviewed material.

After the hypothesis has been generated, an experimental clinical trial is designed to test the

Event Stream Processing

► Complex Event Processing (CEP)

validity of the hypothesis. The clinical trial is designed as a unique study; however, it usually has similarities to other studies that were identified while developing the hypothesis. Before the proposed clinical trial can be performed, it needs to be reviewed and approved by a neutral Institutional Review Board (IRB). The IRB is a group of scientists, practitioners, ethicists, and public representatives who evaluate the research procedures. Members of the IRBs use their expertise to determine if the study is ethical and if the potential benefits of the proposed study will far outweigh its possible risk.

After, and only after, IRB approval has been obtained, the researcher begins the clinical trial by recruiting volunteer participants who of their own free will agree to participate in the research. Participant recruitment is a process whereby a large number of the people who meet certain inclusion criteria (e.g., they have the specific disorder and they are members of the specific population of interest) and who don't have any exclusion criteria (e.g., they don't have some other health condition that might lead to erroneous findings) are identified. These people are then contacted and asked if they would be willing to participate in the study. The risks and benefits of the study are explained to each participant. After a large group of volunteers has been identified, randomization procedures are used to assign each person to one of two different groups. One of the groups is the treatment group; members of this group will all receive the experimental treatment. Members of the other group, the control group, will receive the treatment that is traditionally used to treat the disorder being studied. The randomization process of assigning people to groups helps insure that each participant has an equal chance of appearing in either the treatment or the control group and that the two groups do not differ in some systematic way that might influence the results of the study.

After a reasonable period of time during which the control group received treatment as usual and the treatment group received the experimental treatment, big data techniques are used to analyze

the results of the clinical trial in great detail. This information is big data that reports the characteristics of the people who were in the different groups, the unique experience of each research participant, the proportion of the treatment and the control who got better, the proportion in each group that got worse, the proportion in each group that had no change, the proportion of people in each group that experienced different kinds of side effects, and a description of any unusual or unexpected events that might have occurred during the course of the study.

After data analysis, the researchers prepare an article that gives a detailed description of the logic that they used in designing the clinical trial, the exact and specific methods that were used in conducting the clinical trial, the quantitative and qualitative results of the study, and their interpretation of the results and suggestions for further research. The article is then submitted to the editor of a professional journal. The journal editor identifies several neutral experts in the discipline who are asked to review the article and determine if it has scientific accuracy and value that makes it suitable for publication. These experts judiciously review all aspects of the clinical trial to determine if the proposed new treatment is safe and effective and that, in at least some cases, it is superior to the traditional treatment that is used for the health problem under study. If the experts agree, the article is published in what is called a peer-reviewed health-care journal. The term "peer reviewed" means that an independent and neutral panel of experts has reviewed the article and that this panel believes the article is worthy of dissemination and study by other health-care professionals.

It should be noted that many different studies are usually conducted concerning the same treatment and the same disorder. However, each study is unique in that different people are studied and the treatment may be administered using somewhat different procedures. As an example, the specific people being studied may differ from study to study (e.g., some trials may only include people between the ages of 18 and 35, some

studies may include only Caucasians, some studies may only include people who have had the disease for less than 1 year). Other studies may look at differences in the treatment procedures (e.g., some studies may only use very small doses, other studies may use large doses, some studies may administer the treatment early in the morning, other studies may administer the treatment late at night). Clearly, there are many different variables that can be manipulated, and these changes can affect the outcome of a clinical trial.

After the article has been published, it joins a group of other articles that address the same general topic. Information about the article and all other similar articles is stored in multiple different online journal databases. These journal databases are massive big data files that contain the article citation as well as other important information about the article (e.g., abstract, language, country, key terms). Practitioners, researchers, students, and others consult these journal databases to determine what is known about the effects of different treatments on the health problem being studied. These big data online databases enable users to identify the most current, up-to-date information as well as historical findings about a condition and its treatment. Many journal databases have options that allow users to receive a copy of the full journal article in a matter of seconds; at other times, it may take as long as a week to retrieve an article from some distant country.

Now that the article has been published and listed in online journal databases, it joins a group of articles that all address the same general topic. By using search terms that relate to the specific treatment and the specific condition, an online journal database user can find the published clinical trial article discussed above as well as all of the other articles that concern the same topic. In reviewing the total group of articles, it becomes apparent that some of the articles show that the treatment under study is very highly effective for the condition being investigated, while other studies show it to be less effective. That is, the evidence is highly variable. Meta-analysis is a research technique that is designed to resolve these differences and determine if the new

treatment is in fact an evidence-based treatment. In performing a meta-analysis, researchers use multiple online databases to identify all of the different articles that address the topic of using the specific treatment with the specific disorder. In a meta-analysis, each of the different identified articles is studied in great detail. Then, a statistic called effect size is calculated. The effect size describes the average amount of effect that a specific treatment has on a specific disorder. Each study's effect size is calculated by comparing the amount of improvement in the disorder that occurs when the new treatment is used with the amount of improvement in the disorder when the new treatment is not being used. After the effect size has been calculated for each of the articles that is being reviewed, an average effect size for the new treatment is calculated based on all of the different clinical trials. If the average effect size shows that the treatment has a significant positive effect on the condition being studied, then, and only then, it is labeled an evidence-based treatment, and this information is widely disseminated to practitioners and health researchers throughout the world.

An evidence-based medical practice (EBMP) is one of the ways evidenced-based medicine (EBM) is applied in the delivery of health-care services. As noted earlier in this article, evidence-based medical practice is the practice of medicine and other health services in a way that integrates the health-care provider's expertise, the patient's values, and the best evidence-based medical information. There are many different elements of this type of practice. Some of these are described below.

Automated symptom checkers are online resources that patients can use to help understand and explain their health difficulties. These algorithms are not a substitute for seeking professional help; however, they often give plausible explanations for the patient's objective signs and patient's subjective symptoms. Big data is central for the development of symptom checkers in that they integrate many different types of data from worldwide sources.

Automated appointment schedulers are devices that enable patients to make appointments

with their health-care provider online or by telephone. In large health systems such as a health maintenance organization (HMO), these are big data systems that track the time availability and the location of many different providers. By using these systems, patients are able to schedule their own appointments with a provider at a time and place that best meets their needs. Very often, the automatic appointment scheduler arranges for a reminder postcard, email, and phone calls to the patients to remind them of the time and place of the scheduled appointment.

Electronic health records (EHR) are comprehensive secure electronic files that contain and collate all of the information about all aspects of a patient's health. They contain information that is collected at each outpatient and inpatient health-care encounter. This includes the patient's medical history, all of their diagnoses, all of the medications the patient has ever taken and a list of the medicines the patient is currently taking, all past and current treatment plans, dates and types of all immunizations, allergies, all past radiographic images, and the results from all laboratory and other tests. These data are usually portable, which means that any time a patient sees a provider, the provider can securely access all of the relevant information to provide the best possible care to the patient.

Conclusion

Evidence-based medicine and evidence-based medical practice are recent health-care advances. The use of these mechanisms depends upon the availability of big data, and as they are used, they generate more big data. It is a system that can only lead to improvements in individual's, nation's, and worldwide health improvements.

Cross-References

- [Health Informatics](#)
- [Telemedicine](#)

Further Reading

- De Vreese, L. (2011). Evidence-based medicine and progress in the medical sciences. *Journal of Evaluation in Clinical Practice*, 17(5), 852–856.
- Epstein, I. (2011). Reconciling evidence-based practice, evidence-informed practice, and practice-based research: The role of clinical data-mining. *Social Work*, 56(3), 284–287.
- Ko, M. J., & Lim, T. (2014). Use of big data for evidence-based healthcare. *Journal of the Korean Medical Association*, 2014, 57(5), 413–418.
- Michael, K., & Miller, K. W. (2013). Big data: New opportunities and new challenges. *Computer*, 46(6), 22–24.

Facebook

R. Bruce Anderson^{1,2} and Kassandra Galvez²

¹Earth & Environment, Boston University,
Boston, MA, USA

²Florida Southern College, Lakeland, FL, USA

When it comes to the possibility of gathering truly “big data” of a personal nature, it is impossible to think of a better source than Facebook. Millions use Facebook every day, and likely give up everything they publish to data collectors every time they use it.

Since its 2004 launch, Facebook has become one of the biggest websites in the world with 400 million people visiting the site each month. Facebook allows any person with a valid email address to simply sign up for an account and immediately start connecting with other users known as “friends.” Once you are connected with another “friend,” users are able to view the other person’s information listed such as: birthday, relationship status, and political affiliation; however, some of this information may not listed depending on the various privacy regulations that Facebook has.

By having a Facebook account, users have a main screen known as the “newsfeed.” This “newsfeed” shows users what his or her “friends” are currently doing online either from liking pictures or writing comments on statuses. Additionally, “users” can see what is currently “trending”

meaning what users are currently discussing on Facebook. On Facebook’s main screen, users can go to the trending section and view the top 10 most popular things that all Facebook users are discussing. This feature includes things that just your friends are discussing to what the nation is talking about. Many of these topics are preceded with the number sign, #, to present a hashtag which links all these topics together. All of these conversations, information about individuals, and location materials are potential targets for data harvesting.

These “trending topics” have provided Facebook users instant knowledge about a specific event or topic. Facebook has integrated news with social media. In the last 12 months, traffic from home pages has dropped significantly across many websites while social media’s share of clicks has more than doubled, according to a 2013 review of the BuzzFeed Partner Network, a conglomeration of popular sites including BuzzFeed, the *New York Times*, and Thought Catalog. Facebook, in particular, has opened the spigot, with its outbound links to publishers growing from 62 million to 161 million in 2013. Two years ago, Facebook and Google were equal powers in sending clicks to the BuzzFeed network’s sites. Today Facebook sends 3.5 times more traffic. Facebook has provided its 1 billion users with a new way of accessing the news.

However, such access can have a double edge. For example, during the 2016 US election, hackers from foreign sources apparently took

advantage of the somewhat laissez faire approach Facebook had towards users security (and content control) and set up false accounts to spread fake information about candidates for office.

With the increased use of Facebook, there has also been an increase in research on Facebook. Many psychologists believe that social can add to a child's learning capacity, but it is also associated with a host of psychological disorders. Additionally, social media can be "the training wheels of life" to social networking teens because it allows he or she to post public information onto the sites, see how other users react to that information, and learn as they go. Research has gone as far as to show "that people who engage in more Facebook activities – more status updates, more photo uploads, more "likes" – also display more virtual empathy." An example of this "virtual empathy" is if someone posts he had a difficult day, and you post a comment saying, "Call me if you need anything," this displays virtual empathy in action.

While Facebook has benefited society in a variety of ways, it has also provided a lack of privacy. When creating a Facebook account, you are asked a variety of personal questions that may be shared on your Facebook page. These questions include your: birthday, marital status, political affiliation, phone number, work and educations, family and relationships, and hometown. While some of this private and personal information may not seem as important to some users, it allows other users who are connected to your account to access that information – but access by others is frankly very easy to obtain, making any notion of a firewall between your information and the information of millions of others a very doubtful proposition.

With the introduction of the program "Facebook Pixel" – program sold to advertisers – commercial users can track, "optimize" and retarget site visitors to their products, ads, and related materials – keeping the data they have gathered, of course, for further aggregation. When data such as these are aggregated, projections can be made about mass public consumer

behavior, allowing targeted ad buys, directed "teasers" and the like.

Unfortunately, the release of private information has caused concern. Users of social-networking sites such as Facebook passively accept losing control of their personal information because they are not fully aware – or have given up caring – about the possible implications for their privacy. These users should make sure they understand who could access their profiles, and how that information could be used. They also should familiarize themselves with sites' privacy options.

Facebook has a variety of privacy settings that can be useful to those who are wary of the private information online such as: "public," "friends," "only me," and "custom." Users who elect to have their information "public" allow any person on or off Facebook to see his or her profile and information. The "friends" privacy setting allows users who are connected to another's account to view the profile and information. Additionally, Facebook provides two unique privacy settings which are: "only me" and "custom." The "only me" privacy only allows the account user to view his or her own material. The last privacy setting is "custom" which allows the user to make his own privacy settings which is a two-step process. The first step for this "custom" privacy setting asks users to choose who he or she would like the information to be shared with. The user is able to write "friends" which allows all his or her Facebook friends to view this information or the user can write the names of specific people. If the user decides to opt for the second choice, all the user's information will only be viewed by those specific friends he or she chose. The second step asks users who he or she does not want the information to be shared with. The user can then write the names of users he or she does not want to view the information. While the second step may seem redundant, the option has proven to be useful to teens who do not want their parents to see everything on his or her Facebook page.

While the privacy settings may be handy, there is some information that Facebook makes

publicly available. This type of information consist of the user's name, profile pictures and cover photos, networks, gender, user name, and user ID. The purpose of having your name public is so that other users may be able to search for your account. When searching for a user, profile and cover photos are normally the indicator that he or she has found the person they are searching for. While these photos remain public, the user may change the privacy settings for these specific pictures once you have updated your account and changed them. For example, a user may be tired of the same profile picture so he or she changes it, the old picture would still appear; however, he or she can change that specific picture to private now that there is a new picture in its place. The same can be done with cover photos. Sadly, Facebook does not allow every single picture to be private. Networks allow users to be linked to specific areas such as: college or university, companies, and other organizations. These networks provide other users who have linked to the same network to search for you due to this custom audience. College students will normally link their accounts to these networks so that other college students can search for them with ease. By having this information public, other users can look onto a person's account and recognize the network and make that connection. Lastly, Facebook requires the users gender, username, and user ID to be public. According to Facebook privacy policies, "Gender allows us to refer to you properly." Username and user IDs provide users to supply others with custom links to his or her profile.

When signing up for a Facebook account, many users forget to "read the fine print" meaning that users do not read the terms of agreement which includes Facebook's policy on information. If users were to read this section, he or she may not have created an account in the first place. In this section, there is a sense that Facebook is constantly watching you especially in the section titled "Other Information we receive about you." This section details how and when Facebook is receiving data from his or her account. Facebook states that they receive data whenever you are

using or are running Facebook, such as when you look at another person's account, send or receive a message, search for a "friend" or page, click on view or otherwise interact with things, use a Facebook mobile app, or make purchases through Facebook. Also, not only does Facebook received data when a user posts photos or videos but it also receives data such as the time, date, and place you took the photo or video. The last paragraph of that section states "We only provide data to our advertising partners or customers after we have removed your name and any other personally identifying information from it, or have combined it with other people's data in a way that it no longer personally identifies you." With all the data collected, Facebook customizes the ad's that users see on their account. For example, college student users who constantly view text book websites will start seeing advertisements about these specific websites on his or her Facebook account. Facebook can be compared to as the "all-seeing eye" because it is constantly watching all 20 billion users.

The last section of the Facebook private policy is titled "How we Use the Information We Receive." Facebook states that they use the information received in connection with services and features they provide to users, partners, advertises that purchase ads on the site, and developers of games, applications, and websites users use. With the information received, Facebook uses it as part of their efforts to keep their products, services, and integrations safe and secure; to protect Facebook's or others' rights or property; to provide users with location features and services, to measure and understand the effectiveness of ads shown to users, to suggest Facebook features to make the users experience easier such as: contact importer to find "friends" based on your cell phone contacts; and lastly internal operations that include troubleshooting, data analysis, testing, search and service improvement. According to Facebook, users are granting the use of information by simply signing up for an account. By granting Facebook the use of information, Facebook, in turn, uses the information to provide users with

new and innovative features and services. The Facebook private policy does not state how long it keeps and stores the data for. The only answer users will receive is “we store data for as long as it is necessary to provide products and services to you and others. . . typically, information associated with your account will be kept until your account is deleted.”

Two ways of ridding users of Facebook are deactivation and deletion. Many Facebook users believe that deactivation and deletion are the same; however, these users are mistaken. By deactivating an account, that specific account is put on hold. While other users may not see his or her information, Facebook does not delete any information. Facebook believe that deactivating an account is the same as a user telling Facebook not to delete any information because he or she might want to reactivate the account at some point in the future. The period of deactivation is unlimited—users can deactivate their accounts for years but Facebook will not delete their information. By deleting a Facebook account, the account is permanently deleted from Facebook; however, it takes up to one month to delete but some information may take up to 90 days to erase. The misconception is that the amount and all its information is immediately erased; however, that is not the case, even for Facebook itself.

In the case of resident programs or third parties that access or harvest the data itself, Facebook assumes no responsibility – there is nothing, for example, to force advertisers to give up the mega-data they collect (legally thus far) when consumers visit their site on the platform.

Facebook also provides users with outside applications. These applications allow users to gain a more tailored and unique experience through the social networking site. These applications include games, shopping websites, music players, video streaming websites, organizations, and more. Now in this day and age, websites will ask users to connect their social networking sites to the specific website account. By granting that access, users provide the websites their basic information and email address. Additionally, the website can send the user notifications but more importantly post on your behalf. This means that

the app may post on your Facebook page with any recent activity the user has done through that specific website.

An example of applications in motion is Instagram, which is a photo social networking service owned by Facebook. Users that have both Facebook and Instagram can connect each account for a more personal experience. Instagram users who connect to Facebook accounts provide Facebook with basic information from their Instagram account and in turn, Instagram may access data from Facebook from other pages and applications used. All this connection and data opens up a “Pandora’s box” for users because he or she may not know what application is getting what information and from where. Furthermore, Instagram can access the users data anytime even when he or she may not be using the Instagram application. The application may even post on my behalf which includes objects Instagram users have posted and more. Lastly, each additional application has its own policy terms.

While the nature of people around the world may be to be trusting, users of all ages need to be cautious of what they post online. Recently, there has been a recent shift of job seekers getting asked for Facebook passwords. The reason for this shift is that employers want to view what (possible) future employees are posting. Since the rise of social networking, it has become common for managers to review publicly available Facebook profiles, Twitter accounts and other sites to learn more about job candidates. However, many users, especially on Facebook, have their profiles set to private, making them available only to selected people or certain networks. Other solutions to private Facebook accounts have surfaced such as asking applicants to friend human resource managers or to log in to a company computer during an interview. Once employed, some workers have been required to sign non-disparagement agreements that ban them from talking negatively about an employer on social media. While such measures may seem fair because employers want to have employees that represent themselves well on social media, these policies have raised concerns about the invasion of privacy.

Facebook has been a source for people to connect, communicate, research and follow latest trends, and post; however, it has also become a hub for private information. While Facebook has private policies available for all users to read and understand, users are not particularly interested in those aspects. Unfortunately users, who do not yet understand the importance of private information and create Facebook accounts to communicate with others, do not pay attention to the things he or she is posting. The real question is: is Facebook really upholding their private policy rules and is my information really private? Given the number of pending and active cases against the platform, the answer is likely “not so much.”

In the end, the financial health of such platforms is predicated on turning a tidy profit through the sale of advertising, and advertiser access to the results of the passive and sometimes active harvesting of big data through the site.

Cross-References

- [Cybersecurity](#)
- [Data Mining](#)
- [LinkedIn](#)
- [Social Media](#)

Further Reading

- Brown, U. (2011). The influence of Facebook usage on the academic performance and the quality of life of college students. *Journal of Media and Communication Studies*, 3(4), 144–150. Web.
- Data Use Policy. *Facebook*. 1 Apr 2014. Web. 24 Aug 2014.
- Mcfarland, S. (2012). Job seekers getting asked for Facebook passwords. *USATODAY.com*, 3 Mar 2012. Web 24 Aug 2014.
- Roberts, D., & Kiss, J. (2013). *Twitter, Facebook and more demand sweeping changes to US surveillance*. *Theguardian.com*. Guardian News and Media, 9 Dec 2013. Web 24 Aug 2014.
- Thompson, D. (2014). *The Facebook effect on the news*. *The Atlantic*. Atlantic Media Company, 12 Feb 2014. Web 24 Aug 2014.
- Turgeon, J. (2011). *How Facebook and social media affect the minds of generation next*. *The Huffington Post*. *TheHuffingtonPost.com*, 9 Aug 2011. Web 24 Aug 2014.

Facial Recognition Technologies

Gang Hua

Visual Computing Group, Microsoft Research, Beijing, China

Facial recognition refers to the application of automatically *identifying* or *verifying* a person from face images and videos. Face verification aims at arbitrating if a pair of faces is from the same person or not. While face identification focuses on predicting the identity of a query face given a gallery face dataset with known identities, there are lots of applications of facial recognition technologies, in domains such as security, justice, social networks, military operations, etc.

While early face recognition technologies dealt with face images taken from well-controlled environment, the current focus in facial recognition research is pushing the frontier in handling real-world face images/videos taken from uncontrolled environment. There are two major unconstrained visual sources: (1) face images and videos taken by users and shared on the internet, such as those images uploaded to Facebook, and (2) face videos taken from surveillance cameras.

In these unconstrained domain visual sources, face recognition technologies must contend with uncontrolled lighting, large pose variations, a range of facial expressions, makeup, changes in facial hair, eye-wear, weight gain, aging, and partial occlusions. Recent progresses in real-world face recognition have greatly benefited from the BIG face dataset from unconstrained sources.

History

As one of the most intensively studied areas in computer vision, facial recognition researchers are among the first in advocating systematic data-driven performance benchmarking. We briefly review the history of face recognition research in the context of datasets they evaluated on.

Early work by Woody Bledsoe, Helen Chan Wolf, and Charles Bisson, though not published due to restriction of funded research from an unnamed intelligent agent, could be traced back to 1964 at Panoramic Research, Inc., and later continued by Peter Hart at Stanford Research Institute after 1966. The task was, given a photo, identifying from a book of mug shots, a small set of candidate records that have one matched with the query photo.

Due to the limited capacity of computers back then, human operators were involved in extracting a set of distances among a set of predefined facial landmarks. These distances were then normalized and served as features to match different face photos. The method proposed by Bledsoe was evaluated in a database of 2000 face photos, and it consistently outperformed human judges.

Later work attempted to build fully automated computer program without involving human labors. Takeo Kanade, in 1977, built a benchmark dataset of 20 young people, each with two face images. Hence the dataset consists of 40 face images. Takeo's program conducted fully automated analysis of different facial regions and extracted a set of features to characterize each region. Back then, the evaluation on these 40 digitized face images are considered to be a large-scale evaluation. Takeo also evaluated his algorithm on a dataset of 800 photos later on.

The Eigenfaces approach proposed by Matthew Turk and Alex Pentland in 1991 was the first to have introduced statistical pattern recognition method for face recognition. It conducted principle component analysis (PCA) on a set of face images to identify a subspace representation for face images. The efficacy of the Eigenfaces representation was evaluated on a dataset of 2500 digitized face images.

The Eigenfaces spurred a theme of work, namely, manifold learning, in identifying better face spaces for face recognition, including the Fisherfaces method by Peter Belhumeur et al. in 1997 and the Laplacianfaces method by Shuicheng Yan et al. in 2005, which aims at identifying a discriminative subspace for representing face

images. The sparse representation-based face identification method proposed by John Wright et al. in 2009 can be regarded as a smarter manifold representation for face recognition.

These manifold learning-based face recognition algorithms have largely been evaluated on several popular face recognition benchmarks, including the Yale Face Database, the Extended Yale Face Database B, the ORL dataset, and the PIE dataset. These datasets are in the order of several hundreds to several thousands. One defect of all these subspace representations is that they would fail if the faces are not very well aligned. In other words, they would fail if the faces are under large pose variations.

While these manifold learning-based methods largely operated on raw image pixels, other invariant local descriptor-based method gained its popularity due to their robustness to pose variations. These include the elastic bunch graph matching method in 1999, the local binary pattern-based face recognition method in 2004, the series of local descriptor-based elastic part matching method published by Gang Hua and Haoxiang Li between 2009 and 2015 including the series of probabilistic elastic part (PEP) models published between 2013 and 2015, the Joint Bayesian faces method in 2012, and the Fisher Vector faces in 2013.

The performance of these methods have largely been evaluated on more recent real-world face recognition benchmark dataset including the labeled faces in the wild (LFW), the YouTube Faces Database, and the more recent point-and-shoot dataset. These face datasets are either collected from the Internet or collected by point-and-shoot cameras in unconstrained settings.

Since 2014, we have witnessed a surge of deep learning-based face recognition systems, e.g., the DeepFace system from Facebook, the DeepID systems from the Chinese University of Hong Kong, and the FaceNet system from Google. They are all trained with millions of face images. For example, the DeepFace system from Facebook has leveraged 4.4 million face images from 4030 people from Facebook for training, and

the FaceNet system leveraged 100 million to 200 million faces consistent of about 8M different identities.

In 2014, the US Government funded the JANUS program under the Intelligence Advanced Research Projects Activity (IARPA), which is targeting on pushing the frontiers of facial recognition technology in unconstrained environment and emphasizing the comprehensive modeling of age, pose, illumination, and facial expression (A-PIE) and unifying both image and video face recognition. Accompanied with this program is a face recognition benchmark, namely, IARPA Janus Benchmark, from the National Institute of Standards and Technologies (NIST). The neural aggregation network invented by Gang Hua and his colleagues in 2016 at Microsoft Research is one representative of the current state of the art on this benchmark to date.

Approaches

Face recognition technology can be categorized in different ways. In terms of visual features exploited for face representation and hence for recognition, face recognition algorithms can be categorized as geometric feature based and appearance feature based. While early work has focused on geometric invariants, such as the size of certain facial components, and the distance between certain facial landmarks, modern face recognition algorithms largely focused on modeling the appearances.

From a modeling point of view, facial recognition technologies can be categorized as holistic methods or part-based methods. Holistic methods build the representation based on the holistic appearance of the face. The numerous manifold learning-based methods belong to this category, while part-based methods attempt to characterize each facial part for robust matching. The series of PEP models developed by Gang Hua and Haoxiang Li are one such example.

From the perspective of pattern recognition, face recognition technologies can be categorized

into generative model based and discriminative model based. The seminal Eigenfaces method is a generative model, while the Fisherfaces method is a discriminative model. From 2014, the recent trend in face recognition is to exploit deep neural network to learn discriminative face representations from a large amount of labeled face images.

Datasets and Benchmarks

While early face recognition research worked on proprietary datasets which were not used by other researchers, the face recognition research community is perhaps the earliest in the computer vision community in adopting systematic and public data-driven benchmarking.

This is catalyzed by the FERET dataset funded by US Department of Defense's Counterdrug Technology Development Program through the Defense Advanced Research Projects Agency (DARPA) during 1993 and 1997. The final FERET dataset consists of 14051 8-bit gray-scale images of human heads with views ranging from frontal view to left and right profile. The FERET dataset is the basis of the Face Recognition Vendor Test (FRVT) organized by NIST in 2000 and 2002.

The FRVT in 2006 adopted the face recognition grand challenge (FRGC) dataset, which evaluated performance of facial recognition systems from different vendors on high-resolution still imagery (5–6 megapixels), 3D facial scans, multi-sample still facial imagery, and pre-processing algorithms that compensate for pose and illumination. The winning team is Neven Vision, a Los Angeles start-up. Neven Vision is later acquired by Google.

The most recent FRVT was organized in 2013; facial recognition systems from various vendors are tested to identify up to 1.6 million individuals. The task is largely focused on visa photos and mug shot photos, where the sources are more or less controlled. The system that ranked overall in the top is the NEC system. Other participants include Toshiba, MorphoTrust, Cognitec,

etc. These FRVT tests organized by the US government in the past have largely been focused on more controlled environment. The IAPRA JANUS benchmarking is currently ongoing, which will further stimulate more accurate face recognition technologies.

Meanwhile, there are also widely adopted benchmark dataset from academia, including the early small-scale datasets collected in 1990s, such as the Yale and Extended Yale B datasets, the ORL dataset from the AT&T Labs, and mid-scale datasets such as the PIE and Multi-PIE datasets collected at CMU. These datasets are often collected to evaluate some specific visual variations that confront facial recognition. Specifically, the Yale datasets are designed for modeling illuminations; the ORL dataset is constructed to evaluate occlusion and facial expressions; and the PIE datasets are designed to model poses, illuminations, and facial expressions.

These datasets are more or less taken in well-controlled setting. The labeled faces in the wild (LFW) dataset published in 2007 is the first dataset collected from the Internet, which released publically to the research community, for systematic evaluation of facial recognition technologies in uncontrolled settings. It contains 13,000 images from 5749 celebrities. The benchmark task on LFW has been mainly designed for face verification, with different protocols depending if the algorithms are trained with external data. Later on, the YouTube Faces Database, published in 2011, followed the same protocol as LFW, but each face instance is a video clip instead of a single image. The dataset contains 3425 videos from 1595 people.

One limitation of the LFW dataset as well as the YouTube Faces Database is that the people in these datasets are celebrities. As a result, the photos and videos published are often taken by professional photographers. This is different from photos taken by amateur users in their daily life. This largely motivated the construction of the point-and-shoot dataset, released from the University of Notre Dame. It is composed of 9376 still

images of 293 people and 2802 videos of 256 people. These photos and videos are taken with cheap digital cameras including those on smartphones. Compared with performance of face recognition algorithms on the LFW and YouTube Faces Database, where nearly perfect verification accuracy has been achieved, current state-of-the-art verification accuracy, up to September 2015, on the point-and-shoot dataset, is 58% at the false acceptance rate of 1%, achieved by the team from the Chinese Academy of Science Institute of Computing.

The current largest publically available facial recognition dataset is the MegaFace dataset, with one million faces obtained from Flickr. The current state-of-the-art rank 1 identification accuracy with one million distractors is around 75%.

Software

Well-known commercial software systems that have used facial recognition technology include the Google Picasa photo management system, the Apple iPhoto system, the photo application of Facebook, Windows Live Photo Gallery, Adobe Photoshop Elements, and Sony Picture Motion Browser. The OKAO vision system from Omron provided advanced facial recognition technologies, which has been licensed to various companies for commercial applications.

As software as a service (SaaS) becomes an industry common practice, more and more companies are offering their latest face recognition technologies through the cloud. One of the most matured ones is the Face API provided by Microsoft Cognitive Service. Other similar APIs are also offered by Internet giants such as Baidu and start-ups such as Megvii and SenseTime in China.

Cross-References

- [Biometrics](#)
- [Facebook](#)
- [Social Media](#)

Further Reading

- Belhumeur, P. N., et al. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transcation Pattern Analysis Machine Intelligence*, 19(7), 711–720.
- Chen, D., et al. (2012). Bayesian face revisited: A joint formulation. In *Proceedings of European Conference on Computer Vision*.
- Huang, G. B., et al. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical report (pp. 07–49).
- Kanade, T. (1977). Computer recognition of human faces. *Interdisciplinary Systems Research*, 47.
- Li, H., et al. (2013). Probabilistic elastic matching for pose variant face verification. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Schroff, F., et al. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Simonyan, K., et al. (2014). Fisher vector faces in the wild. In *Proceedings of British Machine Vision Conference*.
- Taigman, Y., et al. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Turk, M. A & Pentland, A. P. (1991) Face recognition using eigenfaces. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Wolf, L., et al. (2011). Face recognition in unconstrained videos with matched background similarity. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Yang, J., et al. Neural aggregation network for video face recognition. <http://arxiv.org/abs/1603.05474>

Factory of the Twenty-First Century

► [Data Center](#)

FAIR Data

► [Data Storage](#)

Financial Data and Trend Prediction

Germán G. Creamer

School of Business, Stevens Institute of Technology, Hoboken, NJ, USA

Synonyms

[Financial econometrics](#); [Machine learning](#); [Pattern recognition](#); [Risk analysis](#); [Time series](#); [Financial forecasting](#)

Introduction

The prediction of financial time series is the primary object of study in the area of financial econometrics. The first step from this perspective is to separate any systematic variation of these series from their random movements. Systematic changes can be caused by trends and seasonal and cyclical variations. Econometric models include different levels of complexity to simulate the existence of these diverse patterns. However, machine-learning algorithms can be used to forecast nonlinear time series as they can learn and evolve jointly with the financial markets.

The most standard econometric approach to forecast trends of financial time series is the Box and Jenkins (1970) methodology. This approach has three major steps: (1) identify the relevant systematic variations of the time series (trend, seasonal or cyclical effects), the input variables, and the dynamic relationship between the input and the target variables; (2) estimate the parameters of the model and the goodness-of-fit statistics of the prediction in relation to the actual data; and (3) forecast the target variable.

The simplest forecasting models are based on either the past values or the error term of the financial time series. The autoregressive model [AR(p)] assumes that the current value depends on the “p” most recent values of the series plus an

error term, while the moving average model [MA(q)] simulates the current values based on the “q” most recent past errors or innovation factors. The combination of these two models leads to the autoregressive moving average [ARMA(p,q)] model, which is the most generic and complete model. These models can include additional features to simulate seasonal or cyclical variations or the effect of external events or variables that may affect the forecast.

Some of the main limitations of this method are that it assumes a linear relationship between the different features when that relationship might be nonlinear, and can manage only a limited number of quantitative variables. Because the complexities of the financial world have grown dramatically in the twenty-first century, better ways of forecasting time series are needed. For example, the 2007–2009 financial crisis brought on a global recession with lingering effects still being felt today. At that point, widespread failures in risk management and corporate governance at almost all the major financial institutions threatened a systemic collapse of the global financial markets leading to large bailouts by governments around the world. Often, the use of simplified forecasting methods was blamed for the lack of transparency of the real risks embedded in the financial assets, and their inability to deal with the complexity of high-frequency datasets.

The high-frequency financial datasets share the four dimensions of big data: an increase in the **volume** of transactions; the high **velocity** of the trades; the **variety** of information, such as text, images and numbers, used in every operation; and the **veracity** of the information in terms of quality and consistency required by the regulators. This explosion of big and nonlinear datasets requires the use of machine-learning algorithms, such as the methods that we introduce in the next sections, that can learn by themselves the changing patterns of the financial markets, and that can combine many different and large datasets.

Technical Analysis

Technical analysis seeks to detect and interpret patterns in past security prices based on charts or

numerical calculations to make investment decisions. Additionally, technical analysis helps to formalize traders’ rules: “buy when it breaks through a high,” “sell when it is declining,” etc. The presence of technical analysis has been very limited in the finance literature because of its lack of a robust statistical or mathematical foundation, its highly subjective nature, and its visual character. In the 1960s and 1970s, researchers studied trading rules based on technical indicators and did not find them profitable. Part of the problem of these studies was the ad hoc specifications of the trading rules that led to data snooping. Later on, Allen and Karjalainen (1999) found profitable trading rules using a genetic algorithmic model for the S&P 500 with daily prices from 1928 to 1995. However, these rules were not consistently better than a simple buy-and-hold strategy.

Machine-Learning Algorithms

Currently, the major stock exchanges such as NYSE and NASDAQ have mostly transformed their markets into electronic financial markets. Players in these markets must process large amounts of structured and unstructured data and make instantaneous investment decisions. As a result of these changes, new machine-learning algorithms that can learn and make intelligent decisions have been adapted to manage large, fast, and diverse financial time series. Machine-learning techniques help investors and corporations discover inefficiencies in financial markets and recognize new business opportunities or potential corporate problems. These discoveries can be used to make a profit and, in turn, reduce the market inefficiencies. Also, corporations could save a significant amount of resources if they can automate certain corporate finance functions such as planning, risk management, investment, and trading.

There is growing interest in applying machine-learning methods to discover new trading rules or to formulate trading strategies using technical indicators or other forecasting methods. Machine learning shares with technical analysis the emphasis on pattern recognition. The main problem with this approach is that every rule may require a

different set of updated parameters that should be adjusted to every particular challenge. Creamer (2012) proposed a method to calibrate a forecasting model using many indicators with different parameters simultaneously. According to this approach, a machine-learning algorithm characterized by robust feature selection capability, such as boosting (described below), can find an optimal combination of the different parameters for each market. Every time that a model is built, its parameters are optimized. This method has shown to be profitable with stocks and futures.

The advantage of machine-learning methods over methods proposed by classical statistics is that they do not estimate the parameters of the underlying distribution and instead focus on making accurate predictions for some variables given others variables. Breiman (2001) contrasts these two approaches as the data modeling culture and the algorithmic modeling culture. While many statisticians adhere to the data-modeling approach, people in other fields of science and engineering use algorithmic modeling to construct predictors with superior accuracy.

Classification of Algorithms to Detect Different Financial Patterns

The following categories describe the application of machine-learning algorithms to various financial forecasting problems:

- Supervised:
 - Classification: Classify observations using two or more categories. These types of algorithms can be beneficial to forecast asset price trends (positive or negative) or to predict customers who may default on their loans or commit fraud. These algorithms can also be used to calculate investors' or news' sentiment.
 - Regression: Forecast future prices or evaluate the effect of several features into the target variable. Results could be similar to those generated by an ARMA model, although machine-learning methods may capture nonlinear relationships among the different variables.
- Unsupervised:
 - Clustering: Aggregate stocks according to their returns and risk to build a diversified portfolio. These techniques can also be used in risk management to segment customers by their risk profile.
 - Modeling: Uncover linear and nonlinear relationships among economic and financial variables.
 - Feature selection: Select the most relevant variables among vast and unstructured datasets that include text, news, and financial variables.
 - Anomaly detection: Identify outliers that may represent a high level of risk. It can also help to build realistic future scenarios to forecast prices, such as anticipating spikes in electricity prices. The complex and chaotic nature of the forces acting on energy and financial markets tend to defeat ARMA models at extreme events.

Learning Algorithms

The following are some of the best well-known learning algorithms that have been used to forecast financial patterns:

Adaboost: Apply a simple learning algorithm to perform an iterative search to locate observations that are difficult to predict, then it generates particular rules to differentiate the most difficult cases. Finally, it classifies every observation combining the different rules generated. This method, invented by Freund and Schapire (1997), has demonstrated to be very useful to support automated trading systems due to its feature selection capability and reliable forecasting potential.

Support vector machine (SVM): Preprocess data in a higher dimension than the original space. As a result of this transformation proposed by Vapnick (1995), observations can be classified into several categories. Support vector machine has been used for feature selection and financial forecasting of several financial products such as the S&P 500 index, and the US and German government bond futures, using moving averages and lagged prices.

C4.5: This is a very popular decision-tree algorithm. It follows a top-down approach where the best feature, introduced as the root of the tree, can separate data according to a test, such as the information gain, and its branches are the values of this feature. This process is repeated successively with the descendants of each node creating new nodes until there are no additional observations. At that point, a leaf node is included with the most common value of the target attribute. Decision trees are very useful to separate customers with different risk profiles. The advantage of decision trees is that their interpretation is very intuitive and may help to detect unknown relationships among the various features.

Neural network (connectionist approach): This is one of the oldest and the most commonly studied algorithms. Most trading systems generate trading rules using neural networks where their primary inputs are technical analysis indicators and the algorithm build different layers of nodes simulating how the brain works. Based on the final result, the model back propagates its errors and corrects the parameters until it has an acceptable accuracy rate. This approach has been applied to forecast and trade S&P 500 index futures, the Warsaw stock price index 20 futures, and Korea stock index 200 futures.

Genetic algorithm (emergent approach): The genetic algorithm or genetic programming approach is used to generate trading rules where its features are coded as evolving chromosomes. These rules are represented as binary trees in which the leaves are technical indicators, and the non-leaves are Boolean functions. Together they represent simple decision functions. The advantage of this approach is that the rules are interpretable and can change according to the financial product under study.

Conclusion

The big data characteristics of the financial data and the modeling requirements allow for very

high parallelism in data stream management, and in the data analysis, either directly or using map/reduce architectures, which in turn will require new algorithms to take full advantage of those characteristics. This will provide some benefits including independent analysis of diverse sources without high, initial synchronization requirements; software that will run on relatively inexpensive, commodity hardware, and a mix of algorithms, along with innovative architectures, that can provide both real-time alerting as well as in-depth analysis.

This paper introduced some of these machine-learning algorithms that can learn new financial markets behaviors, approximate very complex financial patterns embedded in big datasets, and predict trends on financial time series.

Further Reading

- Allen, F., & Karjalainen, R. (1999). Using genetic algorithms to find technical trading rules. *Journal of Financial Economics*, 51(2), 245–271.
- Box, G. Y., Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–215.
- Creamer, G. (2012). Model calibration and automated trading agent for euro futures. *Quantitative Finance*, 12(4), 531–545.
- Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.
- Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Financial Econometrics

- [Financial Data and Trend Prediction](#)

Financial Forecasting

- [Financial Data and Trend Prediction](#)

Financial Services

Paul Anthony Laux

Lerner College of Business and Economics and
J.P. Morgan Chase Fellow, Institute for Financial
Services Analytics, University of Delaware,
Newark, DE, USA

The Nature of the Financial Services Sector

The financial services sector performs functions that are crucial for modern economies. Most centrally these are:

- The transfer of savings from household savers and business investors in capital goods; this enables capital formation and growth in the economy over time.
- The provision of payment systems for good and services; this enables larger, faster, and more cost-efficient markets for goods and services as any point in time.
- The management of risk, including insurance, information, and diversification; this enables individuals and firms to bear average risks and avoid being paralyzed by undue volatility.

These services are provided in contract-like ways (via bank accounts, money market and mutual funds, pensions and 401-Ks, credit cards, business, car and mortgage loans, and life and casualty insurance policies) and in security-like ways (via stocks, bonds, options, futures, and the like). These services are provided by traditional firms (like commercial and investment banks, mutual fund companies, asset managers, insurance companies, and brokerage firms) and new economy enterprises (like peer-to-peer lending, cooperative payment transfer systems, and risk sharing cooperatives). These services are provided by long-standing firms with famous names (like Goldman Sachs, Bank of America, General Electric, Western Union, AIG, and Prudential

Insurance), new or fairly new firms with fairly famous names (like Ally Bank, Fidelity, and Pay Pal), and firms so new and small that only their customers know their names.

The economic importance of the financial services sector can be sensed from its size. For example, financial services accounts for a bit less than 10% of the total value added and GDP of the US economy. It employs around 6 million people in the USA. The US Department of Labor Statistics refers to it as the “Financial Activities Super-sector.” Even further, because its functions are so central to the functioning of the rest of the economy, the financial sector has importance beyond its size. Famously, financial markets and the provision of credit are subject to bouts of instability with serious implications for the economy as a whole.

The Importance of Big Data for the Financial Services Sector

Big data is a fast developing area. This entry focuses on recent developments. A historical discussion of related issues, grouped under the term “e-finance,” is provided by Allen et al. (2002). Another discussion, using the term “electronic finance,” is given by Claessens, Glaessner, and Klingebiel (2002). Even though an attempt to delineate all the connections of big data with financial services runs the risk of being incomplete and soon obsolete, there are some key linkages that appear likely to persist over time. These include:

- Customer analytics and credit assessment
- Fintech
- Financial data security/cybersecurity
- Financial data systems for systemic risk management
- Financial data systems for trading, clearing, and risk management
- Financial modeling and pricing
- Privacy
- Competitive issues

Customer analytics. One of the major early uses of big data methods within the financial services sector is for customer analytics. Several particular uses are especially important. The first of these is the use of big data methods to assess creditworthiness. Banks, credit card issuers, and finance companies need to decide on credit terms for new applicants and, in the case of revolving credit, to periodically update terms for existing customers. The use of payment and credit history, job, and personal characteristics for this purpose is long-ago well developed. However, the use of data acquired as a natural part of doing business with retail credit customers (for example, the purchasing details of credit card users) is only recently being undertaken. In less developed credit markets (China, for example) where credit score and the like are not available, lenders have been experimenting successfully with inferring creditworthiness from Internet purchasing history.

A second prominent use of big data for customer analytics is to tailor the offering of financial products via cross-selling, for example, using credit card purchase data to help decide which insurance products might be attractive.

A third prominent use of big data in financial services customer analytics arises because a credit relationship is inherently long-lived. Thus, decisions must be made over time to tailor the relationship with the customer. A specific example is the use of data extracted from recordings of phone calls with delinquent-pay mortgage borrowers. It is of interest to the lender to distinguish between clients who want to stay in their homes but are experiencing financial trouble from those who are less committed to retaining ownership of a house. Experiments are underway at one large mortgage services provider to analyze voice data (stress levels, word choice, pacing of conversation, etc.) in an attempt to discern the difference. From a broader point of view, many of these activities fit in with recent thinking about customer analytics in service businesses more generally, in that they focus on ways to create value via customer engagement over time, as discussed in, for example, Bijmolt et al. (2010).

Fintech. Fintech refers to the provision of financing and financial services by nontraditional

firms and networks supported by Internet communication and data provision. From a finance point of view, much of fintech fits into “shadow banking,” the provision by nonbanks of financial services traditional provided by banks. Much of fintech also fits into the more general concept of peer-to-peer service provision. Examples of fintech include equity finance for projects via such networks as Kickstarter, interpersonal lending via networks such as Lending Club and Prosper, single currency payment networks like PayPal and CashEdge, and foreign exchange transfer services such as CurrencyFair and Xoom. The operations of many of these services are Internet-enabled but not big-data. Even so, the aggregation and policing of working-size peer-to-peer network definitely involves big-data methods. For more on peer-to-peer payments developments, see Windh (2011). For a broader recent discussion, see *The Economist* (2013). The use of the term “fintech” has evolved over time; for an early commercial use of the term, which also gives a sense for former nature of “big data,” see Bettinger (1972).

Financial data security/cybersecurity. Cybersecurity is a huge and growing need when it comes to financial services, as demonstrated by the frequency and size of successful hacker attacks on financial institutions (such as JP Morgan Chase) and financial transactions at retail firms (such as credit card transactions at Target). The banking system has a long experience of dealing with security in relatively closed systems (such as ATM networks). Increasingly, nonbank firms and more open networks are involved. With the advent of mobile banking, mobile payments, and highly dispersed point of payment networks, the issue will continue to grow in importance.

Financial data systems for systemic risk management. The global financial crisis of 2007–2008, and the recession that followed, exemplified with painful clarity the deep interconnections between the financial system and the real economy. Few expected that consumers’ and banks’ shared incentives to over-leverage real estate investments, the securitization boom that this supported, and the eventual sharp decline in

values could trigger such massive difficulties. These included a near-failure of the banking system, a freeze-up in short-term lending markets, a global bear market in stocks, and extensive and persistent unemployment in the real economy. A core problem was convergence: in a crisis, all markets move down together, even though they may be less-correlated in normal times. Big data research methods have the potential to help reduce the chance of a repeat, by helping us understand the sources of cross-market correlation. For a broader discussion and implications for the future of data systems and risk in the financial system, see Kyle et al. (2013).

Financial data systems for trading, clearing, and risk management. From an economic point of view, banks, financial markets, businesses, and consumers are intrinsically interconnected, with effects flowing from one to the others in a global system. From a data systems point of view, the picture is more of moats than rivers. For example, the United States uses one numbering system for tracking stocks and bonds (CUSIP), while the rest of the world is not on this standard. Within the United States, the numbering systems for mortgage loans and mortgage securities are not tightly standardized with those for other financial products. To systematically trace from, say, payroll data to purchase of goods and services via credit cards, to effects on ability of make mortgage payments, to the default experience on mortgage bonds is, to put it lightly, hugely difficult. The development of ontologies (structural frameworks for systematically categorizing and linking information) is a growing need.

Financial modeling and pricing. One of the earliest uses of extensive computing power within the banking industry was for large simulation models that could inform the buying, selling, and risk management of loans, bonds, derivatives, and other securities and contracts. This activity has become extremely well developed and somewhat commoditized at this point. Big data, in the sense of unstructured data, has been less used, though cutting-edge computer scientific methods are routinely employed. In particular, neural network methods (less recently) and machine

learning methods (more recently) have been explored with particular application to trading and portfolio management.

Privacy. The tradeoffs of privacy and benefit for big data in financial services are qualitatively similar to those in other sectors. However, the issues are more central given that personal and company information is at the heart of the sector's work. With more amassing, sharing, and analysis of data collected for one purpose to serve another purpose, there can be more benefit for the consumer/client and/or for the firm providing the service. Conflicts of interest are numerous, and the temptation (or even tendency) will be to use the consumers' data for the benefit of the firm, or for the benefit of a third party to which the data is sold. As in other sectors, establishing clear ownership of the data themselves is key, as is establishing guidelines and legal limits for their use. Privacy issues seem certain to be central to a continuing public policy debate and to *Competitive issues*. Just as many of the big-data uses listed above have privacy implications, they also have implications for financial firms' competitiveness. That is, big data may help us to better understand the interconnections of consumers, housing, banks, and financial markets if we can link consumer purchases, mortgage payment, borrowing, and stock trading. But financial products and services are relatively easy to duplicate, so customer identities and relationships are a special and often secret asset. In finance, information is a competitive edge and is likely to be jealously guarded.

Further Reading

- Allen, F., McAndrews, J., & Strahan, P. (2002). E-finance: An introduction. *Journal of Financial Services Research*, 22(1–2), 5–27.
- Bettinger, A. (1972). FINTECH: A series of 40 time shared models used at manufacturers Hanover trust company. *Interfaces*, 2, 62–63.
- Bijmolt, T. H., Leeflang, P. S., Block, F., Eisenbeiss, M., Hardie, B. G., Lemmens, A., & Saffert, P. (2010). Analytics for customer engagement. *Journal of Service Research*, 13(3), 341–356.
- Kyle, A., Raschid, L., & Jagadish, L.V. (2013). *Next generation community financial cyberinfrastructure for*

managing systemic risk, National Science Foundation Report for Grant IIS1237476.

The Economist. (2013). Revenge of the nerds: Financial-technology firms, 03 Aug, 408, 59.

Windh, J. (2011). *Peer-to-peer payments: Surveying a rapidly changing landscape*, Federal Reserve Bank of Atlanta, 15 Aug.

Forester

► Forestry

Forestry

Christopher Round

George Mason University, Fairfax, VA, USA

Booz Allen Hamilton, Inc., McLean, VA, USA

Synonyms

Forester; Silviculture; Verderer

Definition

Forestry is the science or practice of planting, managing, and caring for forests to meet human goals and environmental benefits (Merriam-Webster 2019). While originally viewed as a separate science, today it is considered a land-use science similar to agriculture. Someone who performs forestry is a forester. Forestry as a discipline pulls from the fields of environmental science, ecology, and genetics. Forestry can be applied to a myriad of goals such as but not limited to timber management for resource extraction, long-term forest management for carbon sequestration, and ecosystem management to achieve conservation goals. Forestry is ultimately a data-driven practice, relying on a combination of the previous experience of the forester and growth models. Big data is increasingly important for the field of forestry as it can improve both the knowledge and process of supply chain management, optimum

growth and harvest strategies, and how to optimize forest management for different goals.

What Is Forestry?

Forestry is the science or practice of planting, managing, and caring for forests to meet human goals and environmental benefits (Merriam-Webster 2019). As a field, forestry has a long history, with evidence of practices dating back to ancient times (Pope et al. 2018). While originally viewed as a separate science, today it is considered a land-use science similar to agriculture. Someone who performs forestry is a forester. Forestry as a discipline pulls from the fields of environmental science, ecology, and genetics (Pope et al. 2018).

Relations to Other Disciplines

Forestry is differentiated from forest ecology and that forest ecology is a value neutral study of forests as ecosystems (Barnes et al. 1998; Pope et al. 2018). Forestry is not value neutral as it is focused on studying how to use forest ecosystems to achieve different socioeconomic and/or environmental conservation goals. While natural resource management, which is focused on the long-term management of natural resources over often intergenerational time scales, may use techniques from forestry, forestry is a distinct discipline (Epstein 2016).

Forestry is related to silviculture, and the two terms have been used interchangeably (Pope et al. 2018; United States Forest Service 2018). Silviculture however is exclusively concerned with growth, composition, and the establishment of timber (Pope et al. 2018), while forestry has a broader focus on the forest ecosystem. Thus, silviculture can be considered a subset of forestry.

Types of Forestry

Forestry can be applied to a myriad of goals such as but not limited to timber management for resource extraction, long-term forest management

for carbon sequestration, and ecosystem management to achieve conservation goals. Modern forestry is focused on the idea of forests having multiple uses (also known as the multiple-use concept) (Barnes et al. 1998; Pope et al. 2018; Timsina 2003). This leads to focuses on sustained yield of forest products as well as recreational activities and wildlife conservation. Sustained yield is act of extracting ecological resources without reducing the base of the resources themselves. This is to avoid the loss of ecological resources. Forestry can be used to manage watersheds and prevent issues with erosion (Pope et al. 2018). It is also connected to fire prevention, insect and disease control, and urban forestry (forestry in urban settings). Urban forestry is of particular concern for the burgeoning field of urban ecology (Francis and Chadwick 2013; Savard et al. 2000; United States Forest Service 2018).

Examples of Prominent Journals

Journal of Forestry published by the Society of American Foresters (Scimago Institutions Rankings 2018)

Forest Ecology and Management published by Elsevier BV (Scimago Institutions Rankings 2018)

Forestry published by Oxford University Press (Scimago Institutions Rankings 2018)

Further Reading

Barnes, B. V., Zak, D. R., Denton, S. R., & Spurr, S. H. (1998). *Forest ecology* (4th ed.). New York: Wiley.

Epstein, C. (2016). Natural resource management. Retrieved 28 July 2018, from <https://www.britannica.com/topic/natural-resource-management>

Francis, R., & Chadwick, M. (2013). *Urban ecosystems: Understanding the human environment*. New York: Routledge.

Merriam-Webster. (2019). Definition of FORESTRY. Retrieved 12, September 2019, from <https://www.merriam-webster.com/dictionary/forestry>

Pope, P. E., Chaney, W. R., & Edlin, H. L. (2018, June 14). Forestry – Purposes and techniques of forest management. Retrieved 25 July 2018, from <https://www.britannica.com/science/forestry>

Savard, J.-P. L., Clergeau, P., & Mennechez, G. (2000). Biodiversity concepts and urban ecosystems. *Landscape and Urban Planning*, 48(3–4), 131–142. [https://doi.org/10.1016/S0169-2046\(00\)00037-2](https://doi.org/10.1016/S0169-2046(00)00037-2).

Scimago Institutions Rankings. (2018). *Journal Rankings on Forestry*. Retrieved 28 July 2018, from <https://www.scimagojr.com/journalrank.php?category=1107>

Timsina, N. P. (2003). Promoting social justice and conserving montane forest environments: A case study of Nepal's community forestry programme. *The Geographical Journal*, 169(3), 236–242.

United States Forest Service. (2018). Silviculture. Retrieved 28 July 2018, from <https://www.fs.fed.us/forestmanagement/vegetation-management/silviculture/index.shtml>

Fourth Amendment

Dzmitry Yuran

School of Arts and Communication, Florida
Institute of Technology, Melbourne, FL, USA

The Fourth Amendment to the US Constitution is fundamental for the privacy law. Part of the US Bill of Rights, ratified in 1791 and adopted in 1792, it was designed to ensure protection for citizens against unlawful and unreasonable searches and seizures of property by the government. The prime role of the Fourth Amendment has not changed since the eighteenth century, but today's expanded number of threats to citizens' privacy demands a wider range of applications for the amendment and brings to life a number of necessary clarifications.

Great amounts of information are generated by organizations and individuals every day. Ever-evolving technology makes capturing and storing this information increasingly simple by turning it into an automated, relatively cheap routine every office and private business owner, website administrator and blogger, smartphone user and video gamer, car driver and most anyone else engage in every day. The ease of storing, accessing, analyzing, and transferring digital information, made possible by technological advances, creates additional vulnerabilities to citizens' privacy and security. Laws and statutes have been put in place to protect privacy of US citizens and shield them

from governmental and corporate abuse. Many argue these regulations struggle to keep up with the rapidly evolving world of digital communications and are turning into obsolete and largely meaningless legislation. While clarifying certain moments about the ways in which US citizens' privacy is protected by the law, the statutes ratified by the US government in its struggle to ensure safety of the nation curb the protective power of constitutional privacy law. And while constitutional law limits, to certain degree, the ability of law enforcement agencies and other governmental bodies to gather and use data on US citizens, private companies and contractors collect information that later could be used by the government and other entities, raising additional concerns.

Like many articles of the US Constitution and of the Bill of Rights, the Fourth Amendment has undergone a great deal of interpretations in court decisions and has been supplemented and limited by acts and statutes enacted by various bodies of the US government. In order to understand its applications in the modern environment, one has to consider the ever-evolving legal context as well as the evolution of the areas of application of the law. The latter is highly affected by accelerating development of technology and is changing too quickly, some argue, for the law to keep up.

The full text of the Fourth Amendment was drafted in the late eighteenth century, during the times when privacy concerns had not yet become topical. As the urban population of the United States was not much larger than 10%, no declarations of privacy were pertinent outdoors and in shared quarters of one-room farm houses.

Amendment IV

The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

Before the electrical telegraph came into existence and well before Facebook started using

personal information for targeted advertisement and the PRISM program was an issue, the initial amendment text was mainly concerned with physical spaces of citizens and their physical possessions. Up until the late 1960s, interpretations of the Fourth Amendment did not consider electronic surveillance without physical invasion of a protected area to be a violation of one's constitutional rights. In other words, wiretapping and intercepting communications were legal. The Supreme Court decision in *Katz v. United States* (1967) case (in which attaching listening and recording devices on telephone booths in public places was contended a violation of the Fourth Amendment) signified recognition of the existence of a constitutional right for privacy and expanded the Fourth Amendment protection from places onto people.

Privacy is a constitutional right, despite the fact that the word "privacy" is not mentioned anywhere in the Constitution. That applies to the Bill of Rights and its first 8 and the 14th amendments, which combined are used to justify the right to be let alone by the government. The US Supreme Court decisions (beginning with the 1961 *Mapp v. Ohio* search and seizure case) created a precedent protecting citizens against unwarranted intrusions by the police and the government. The ruling in *Katz* added one's electronic communications to the list of private things and extended private space into public areas. However, legislation was enacted by congress since, which broadened significantly the authority of law enforcement agencies in conducting surveillance of citizens.

Constitutional law is not the only regulation of electronic surveillance and other forms of privacy infringement – federal laws as well as state statutes play crucial parts in the process as well. While individual state legislation acts vary significantly and have limited areas of influence, a rather complex statutory scheme implemented by Congress applies across state borders and law enforcement agencies. The below four statutes were designed to supplement and specify application of the constitutional right to be let alone in specific circumstances. At the same time, they open up new opportunities for law enforcement

and national security services to collect and analyze information about private citizens.

Title III of the Omnibus Crime Control and Safe Streets Act (Title III) of 1968 regulates authorized electronic surveillance and wiretapping. It requires law enforcement representatives to obtain a court order prior to intercepting private citizens' communications. If state legislation does not allow for issuance of such orders, wiretapping and other forms of electronic surveillance could not be authorized and carried out legally. The regulated surveillance includes that to which neither of the parties in the surveyed communication gives their consent – in other words, eavesdropping. “National security” eavesdropping has an exceptional status under the statute.

The Electronic Communications and Privacy Act (EPCA) of 1986 amended Title III and added emails, voicemail, computing services, and wireless telephony to the list of regulated communications. And while the purpose of EPCA was to safeguard electronic communications from government intrusion and to keep electronic service providers from accessing personal information without users' consent, the statute did add to the powers of law enforcement in electronic surveillance in some circumstances.

The Communications Assistance for Law Enforcement Act (CALEA) of 1994, in the language of the act, made “clear a telecommunications carrier's duty to cooperate in the interception of communications for law enforcement purposes.” The statute was also aimed at safeguarding law enforcement's authorized surveillance, protecting privacy of citizens and protecting the development of new technology. Yet again, while more clarity was brought to protecting citizens' rights in light of new technological advances, more authority was granted to law enforcement agencies in monitoring and accessing electronic communications.

The heaviest blow that constitutional privacy rights had suffered from congressional statutes was dealt by the USA PATRIOT Act. Passed after the terrorist attacks on the United States on September 11, 2001, it was designed to address issues with communications between and within

governmental agencies. While data sharing was streamlined and regulated, the investigative powers of several agencies were significantly increased.

The dualistic nature of the four statutes mentioned above makes them a source of great controversy in the legal and the political worlds. On the one hand, all four contain claims of protecting individuals' privacy. On the other hand, they extend the law enforcement agencies' power and authority, which in return limits rights of individuals. The proponents of the statutes argue that empowering law enforcement is absolutely necessary in combating terrorism and fighting crime, while their opponents raise the concern of undermining the Constitution and violating individual rights.

While the main scope of constitutional law is on protecting citizens from governmental abuses, regulation of data gathering and storage by private entities has proven to be altogether critical in this regard as well. Governmental entities and agencies outsource some of their operations to private contractors. According to the Office of the Director of National Intelligence, nearly 20% of the intelligence personnel worked in private sector in 2013. Bloomberg Industries analyses showed about 70% of the intelligence budget going to contractors that year. While the intelligence agencies do not have sufficient resources to oversee all the outsourced operations, access of private entities to sensitive information brings with it risks to national security. Disclosure of some of that information has also revealed questionable operations by the government agencies, like the extent of National Security Agency's eavesdropping, leaked by Edward Snowden, former employee of Booz Allen Hamilton consulting firm. Some argue that the latter is a positive stimulus for the development of transparency, which is crucial for healthy evolution of a democratic country. The other side of the argument focuses attention on the threats to national security brought about by untimely disclosure of secret governmental information.

Besides commissioning intelligence work to the private sector, the government often requests information from organizations unaffiliated with

the federation or individual states. According to Google Transparency Report, during the 2013 calendar year, the corporation received 21,492 requests (over 40% of all requests by all countries that year) for information about its 39,937 users from the US government and provided some data on 83% of the requests. While not having to maintain files on private citizens and engage in surveillance activities requiring court orders, the government can get access to vast amounts of data collected by private entities for their own purposes. What kind of data Google and other private companies can store and then provide to the US government upon legitimate congressional statutes acts requests is just as much of a Fourth Amendment issue and a government access to information concern as NSA eavesdropping.

In the 1979 *Miller vs. Maryland* case, the Supreme Court of the United States ruled that willingly disclosed information was not protected by the Fourth Amendment. Advances in technology allow for using the broad “voluntary exposure” terminology for a wide variety of information gathering techniques.

Quite a few things users share with electronic communications service providers voluntarily and knowingly. In many cases, however, people are unaware of the purposes that the data collected from them are serving. Sometimes, they don’t even know that information is being collected at all. As an example, Google installed tracking cookies bypassing Safari’s (Web browser developed by Apple Inc.) privacy settings and had to pay a substantial fine (nearly 23 million dollars) for such behavior. The case illustrates that companies would go to great lengths and will engage in questionable activities to gather more information, often in violation of users’ expectations and without users possibly foreseeing it. As another example, Twitter, a social networking giant, used its iPhone application software to upload and store all the email addresses and phone numbers from its users’ devices in 2012 while leaving service subscribers oblivious to the move. The company kept the data for 18 months. In another instance from the same year, Walmart bought Social Calendar, a Facebook application. At the time, the

app’s database contained information about 15 million users, including 110 million birthdays and other events. Users provided this information in order to receive updates about their family and friends, and now it ended up in hands of a publicly held company, free to do anything they would like with that data.

Collecting data about users’ activities on the Internet via tracking cookies is considered voluntary exposure, even though users are completely unaware of what exactly is being tracked. No affirmative consent is needed for keeping log of what a person reads, buys, or watches.

According to a 2008 study out of Carnegie Mellon University, it would take an average person 250 working hours a year to read all the privacy policy statements for all the new websites they visited, more than 30 full working days. By 2016, 8 years later, the number has likely increased considering the rise in the time people spend online. Not surprisingly, the percentage of people who actually read privacy statements of the websites they visit is very low. And while privacy settings create an illusion of freedom from intrusion for the users of social networking sites and cloud-storage services alike, they grant no legal protection.

As users of digital services provide their information for the purpose of getting social network updates, gift recommendations, trip advice, purchase discounts, etc., it often ends up stored, sold, and used elsewhere, often for marketing purposes and sometimes by the government.

Storage and manipulation of certain information pose risks to the privacy and safety of people providing it. By compiling bits and pieces of information, one can establish an identifiable profile. In an illustrative example, America Online released 20 million Web searches over a three-month period by 650,000 users. While no names or IP addresses were left in the data set, the New York Times managed to piece together profiles from the data, full enough to establish people’s identities. And while a private company has gone so far just for the sake of proving the concept, governmental agencies, foreign regimes, terrorist organizations, and criminals could

technically do the same with very different goals in mind.

Constitutional law of privacy has been evolving alongside with technological developments and state security challenges. Due to the complexity of the relationships between the governmental agencies and the private sector today, its reach exceeds prescription of firsthand interactions between law enforcement officers and private citizens. Information gathered and stored by third parties allows for governmental infringement on its citizens privacy just the same. And while vague language of “voluntary disclosure” and “informed consent” is being used by private companies to collect private information, users are often unaware of the uses for the data they provide and potential risks from its disclosure.

Cross-References

- [National Security Administration \(NSA\)](#)
- [Privacy](#)

Further Reading

- Carmen, D., Rolando, V., & Hemmens, C. (2010). *Criminal procedure and the supreme court: A guide to the major decisions on search and seizure, privacy, and individual rights*. Lanham, MD: Rowman & Littlefield Publishers.
- Gray, D., & Citron, D. (2013). The right to quantitative privacy. *Minnesota Law Review*, 98(1), 62–144.
- Joh, E. E. (2014). Policing by numbers: Big data and the fourth amendment. *Washington Law Review*, 89(1), 35–68.
- McInnis, T. N. (2009). *The evolution of the fourth amendment*. Lanham: Lexington books.
- Ness, D. W. (2013). Information overload: Why omnipresent technology and the rise of big data Shouldn't spell the end for privacy as we know it. *Cardozo Arts & Entertainment Law Journal*, 31(3), 925–957.
- Schulhofer, S. J. (2012). *More essential than ever: The fourth amendment in the twenty first century*. Oxford: Oxford University Press.
- United States Courts. What does the fourth amendment mean? <http://www.uscourts.gov/educational-resources/get-involved/constitution-activities/fourth-amendment/fourth-amendment-mean.aspx>. Accessed August, 2014.

Fourth Industrial Revolution

Laurie A. Schintler

George Mason University, Fairfax, VA, USA

Overview

The Fourth Industrial Revolution (4IR) is just beginning to unfold and take shape. Characterized by developments and breakthroughs in an array of emerging technologies (e.g., nanotechnology, artificial intelligence, blockchain, 3D printing, quantum computing, etc.), the 4IR – also known as Industry 4.0 – follows from three prior industrial revolutions (Schwab 2015):

1. Steam power and mechanization of manufacturing and agriculture (eighteenth century)
2. Electricity and mass production (early nineteenth century)
3. Information technology and automation of routine manual and cognitive processes (second half of the twentieth century)

While the 4IR builds on digital technologies and platforms that arose in the Third Industrial Revolution, i.e., the “digital revolution,” this latest period of disruptive technological and social change is distinct and unprecedented in its “velocity, scope, and systems” impact (Schwab 2015). Technology is progressing at an accelerating rate, advancing exponentially rather than linearly. Moreover, technologies tied to the 4IR touch large swaths of the globe and every industry and sector, “transforming entire systems of production, management, and governance” (Schwab 2015). Emerging technologies are also blurring the boundaries between the “physical, digital, and biological” worlds (Schwab 2015), with the capacity to assist, augment, and automate human behavior and intelligence in ways not possible before. Indeed, the 4IR is radically reshaping how we live, work, interact, and play in novel and remarkable ways.

Big data and big data analytics are vital elements of the Fourth Industrial Revolution. They play a prominent and essential role in all the technological pillars of the 4IR, including cyber-physical systems (CPS), the Internet of Things (IoT), cloud computing, artificial intelligence (AI), and blockchain, among others. As critical inputs and outputs to these systems and their components and related applications, big data (big data analytics) can be considered the connective glue of the 4IR.

Role of Big Data and Analytics

Cyber-physical systems (CPSs), which are “smart systems” that integrate physical and cyber components seamlessly and automatically to perform sensing, actuating, computing, and communicating functions for controlling real-world systems, are big data engines (Atat et al. 2018). CPSs are everywhere – from autonomous vehicles to the smart grid to industrial control and robotics systems, and they are contributing to a tsunami of big data (Atat et al. 2018). Consider a single autonomous vehicle, which produces 4,000 gigabytes of data per data for just a single hour of driving (Nelson 2016). To handle the massive amounts of data it generates, a CPS relies on two functional components: system infrastructure and big data analytics (Xu and Duan 2019). The former supports activities tied to data acquisition, storage, and computing, while the latter enables real-time, actionable insight to be gleaned from the data. Both are critical for ensuring that CPSs are scalable, secure, resilient, and efficient and that the products and services they provide are customized to the needs and desires of consumers (Xu and Duan 2019).

The Internet of Things (IoT) serves as a critical bridge between CPSs, enabling data and information exchange between systems (Atat et al. 2018). The IoT is a massive (and continually growing) network of machine devices (e.g., sensors, robots, and wearables) – or “smart objects” – tied to the Internet. Each object has a “unique identifier” and the capacity to transfer data over a network without the need for a human-in-the-loop (Rose et al. 2015). Devices

connected to the IoT and IoT applications (e.g., weather prediction systems, smart cities, and precision agriculture) produce continual data streams, thus providing an ongoing and real-time picture of people, places, industries, and the environment.

Big data generated by CPSs, IoT, and other technologies rely heavily on distributed storage and processing technology based on cloud computing (Hashem et al. 2015). Cloud computing captures, stores, and processes information at data centers in the “cloud,” rather than on a local electronic device such as a computer. Through resource virtualization, parallel processing, and other mechanisms, cloud computing facilitates scalable data operations and warehousing (Hashem et al. 2015). Moreover, given it uses a “software-as-a-service” model, any person, place, or organization can access it – at least, in theory. However, with all that said, cloud computing is not an efficient or scalable solution for managing the deluge of geo-temporal data produced by mobile devices and spatially embedded sensors, such as those connected to the IoT and CPSs. Accordingly, there is a move toward edge computing, which handles data at its source rather than at a centralized server.

Traditional computational and statistical approaches and techniques cannot effectively accommodate and analyze the volume and complexity of data produced by CPSs, the IoT, and other sources (Ochoa et al. 2017). In this regard, artificial intelligence (AI), referring to a suite of data-driven methods that mimic various aspects of human information processing and intelligence, has a critical role to play. For example, deep learning, a particular type of AI, has the capacity for “extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks” (Najafabadi et al. 2015). Cognitive computing is an emerging analytical paradigm, which leverages and integrates various methods and frameworks, including deep learning, natural language processing, and ontologies (Hurwitz et al. 2015). In contrast to AI alone, cognitive computing can learn at scale, interact with reason, understand

the context, and naturally interact with humans (Vajradhar 2019). Therefore, it is a highly intelligent human-centered approach for making sense of and gaining actionable knowledge from big data.

The quality, integrity, and security of big data used and produced by technological systems in the 4IR are enormous concerns. Blockchain, a decentralized, distributed, and immutable digital ledger, provides a possible means for addressing these issues. Unlike centralized ledgers, blockchain records transactions between parties directly, thus removing the intermediary. Each transaction is vetted and authenticated by powerful computer algorithms running across all the blocks and all the users, where consensus across the nodes is required to establish a transaction's legitimacy. All data on the blockchain is encrypted and hashed. Accordingly, any data added to a blockchain should be accurate, immutable, and safe from intrusion and unauthorized alterations; however, in reality, this is not always the case (Alkhalifah et al. 2019).

Conclusion

As with previous industrial revolutions, the 4IR is likely to benefit society in various ways – e.g., by increasing productivity, efficiency, and quality of life. However, it also comes with some downsides and dangers. The use and application of emerging technologies and big data raise various social and ethical issues and challenges. In this regard, one grave concern is that the 4IR could deepen existing gaps and disparities, such as the digital divides, or contribute to new inequalities and inequities altogether. Algorithmic bias and discrimination, privacy infringement, and degradation of human autonomy are additional concerns. Given the dehumanizing effects of the 4IR, some are envisioning the next industrial revolution, i.e., the Fifth Industrial Revolution (5IR), which ideally would facilitate trust in big data and technology by bringing humans back into proper focus (World Economic Forum 2019). In other words, in the 5IR, “humans and machines will dance

together,” ensuring that, ultimately, humanity remains in the loop.

Cross-References

- Artificial Intelligence
- Blockchain
- Internet of Things (IoT)

Further Reading

- Alkhalifah, A., Ng, A., Kayes, A. S. M., Chowdhury, J., Alazab, M., & Watters, P. (2019). A taxonomy of blockchain threats and vulnerabilities. *Preprints*.
- Atat, R., Liu, L., Wu, J., Li, G., Ye, C., & Yang, Y. (2018). Big data meet cyber-physical systems: A panoramic survey. *IEEE Access*, 6, 73603–73636.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115.
- Hurwitz, J., Kaufman, M., Bowles, A., Nugent, A., Kobielus, J. G., & Kowolenko, M. D. (2015). *Cognitive computing and big data analytics*. Indianapolis: Wiley.
- Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 1.
- Nelson, P. (2016). Just one autonomous car will use 4,000 GB of data/day. *Networkworld*. December 7, 2016. <https://www.networkworld.com/article/3147892/one-autonomous-car-will-use-4000-gb-of-dataday.html>. Accessed 1 Feb 2021.
- Ochoa, S. F., Fortino, G., & Di Fatta, G. (2017). Cyber-physical systems, internet of things and big data. *Future Generation Computer Systems*, 75, 82–84.
- Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. *The Internet Society (ISOC)*, 80, 1–50.
- Schwab, K. (2015). The Fourth Industrial Revolution: What It Means and How to Respond. Retrieved from <https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution>
- Vajradhar, V. (2019). Benefits of cognitive technology. *Medium*. November 6, 2019. <https://pvvajradhar.medium.com/benefits-of-cognitive-technology-c1bf35e4b103>. Accessed 19 Jan 2021.
- World Economic Forum. (2019). What the Fifth Industrial Revolution is and why it matters. <https://europeansting.com/2019/05/16/what-the-fifth-industrial-revolution-is-and-why-it-matters/>. Accessed 18 Jan 2021.
- Xu, L. D., & Duan, L. (2019). Big data for cyber physical systems in industry 4.0: A survey. *Enterprise Information Systems*, 13(2), 148–169.

Fourth Paradigm

Kristin M. Tolle

University of Washington, eScience Institute,
Redmond, WA, USA

In 2009 when the book *The Fourth Paradigm: Data-intensive Scientific Discovery* (Hey et al. 2009) was published, few people understood the impact on science we are experiencing today as a result of having an overwhelming prevalence of data. To cope with this situation, the book espoused a shift to multidisciplinary research. The new reality of science was that scientists would either need to develop big data computing skills or, more likely, collaborate with computing experts to shorten time to discovery. It also proposed that data, computing, and science, **combined**, would facilitate a future of fundamental and amazing discovery.

Rather than heralding the obsolescence of scientific methods as some have suggested (Anderson 2007), *The Fourth Paradigm* espoused that science, big data, and technology, **together**, were greater than the sum of their parts. One cannot have science without scientists or scientific methodology. This paradigm shift was for **science, as a whole** (well beyond the examples provided in the book). The point was that with the advancement of data generation and capture, science would need additional technological and educational advances to cope with the coming data deluge.

The extent of the fourth paradigm shift is as defined by physicist Thomas Kuhn in his 1962 book, *The Structure of Scientific Revolutions* (Kuhn 1962): “a fundamental change in the basic concepts and experimental practices of a scientific discipline.” To establish a baseline for the fourth paradigm and beyond, it is important to recognize the three earlier scientific paradigms: observation, modeling, and simulation. The availability of big data is the fourth and, as of this writing, current scientific paradigm shift.

The application of each of these scientific methodologies had profound impacts on science, and all are in use today. The availability of big

data does not mean scientists no longer need to collect small data observations. In fact, each of the paradigm shifts, as they sequentially emerged, were in many ways complementary methodologies to facilitate scientific discovery.

A current example of an ongoing scientific endeavor that employs all four paradigms is NASA’s mission to Jupiter with the Juno spacecraft (NASA 2020). Simulation (Third Paradigm) is critical to scientific data collection when exact replication is not possible regarding the conditions under which these data will be collected. Simulation is valuable to extrapolate how models developed on actual observations can potentially apply to other celestial objects, both intra- and extra-solar. The spacecraft will not only be collecting massive amounts of streaming data (Fourth Paradigm), it will also collect individual direct observations (First Paradigm) such as when it plunges into Jupiter’s atmosphere. Modeling using statistical and machine learning techniques (Second Paradigm) will be needed to extrapolate beyond observations to create knowledge from data.

Various conclusions can be drawn from this example. (1) No paradigm obviates the need for the other. (2) Machine learning and artificial intelligence are comprehensively covered by the second paradigm. (3) It is virtually impossible for one person to be an expert at all four paradigms – thus, the need for collaboration. This last point bears further clarification. Experts in chemistry need not be experts in cloud computing and AI. In fact, their depth of knowledge acts as a method to evaluate all four paradigms. They have the ability to see a flaw in a computer simulation that a modeler often cannot, and though, over time, an applied mathematician might get better at seeing simulation anomalies, there is always a need for someone with extensive knowledge of a scientific discipline as a fail safe to drawing inaccurate conclusions and potentially determine that some data collections may be invalid or not applicable.

Much has changed for scientists since *The Fourth Paradigm* was published in 2009. A crucial change was the recognition and formalization of the field of data science as an educational discipline. Due to a dearth of professionals

needed in this area, the National Academies of Sciences, Engineering, and Medicine in the United States (U.S.) convened a committee and developed a report guiding the development of undergraduate programs for higher education with the goal of developing a consistent curriculum across undergraduate institutions (National Academies of Sciences, Engineering, and Medicine 2018). Also impacting education is the availability of online education and, today, the risks involved with in-person education in the face of the COVID-19 pandemic (Q&A with Paul Reville 2020).

This is not to say that data science is a “new” field of study. For example, a case can be made that the “father” of modern genetics, Johann Mendel [1822–1884], used data science in his genetics experimentation with peas. Earlier still, Nicolaus Copernicus [1473–1543] used data observations and analysis to determine that the planets of our solar system orbited the sun, not the Earth as espoused by the educators of the day. Their data were “small” and analysis methodologies simplistic, but their impact on science, driven by data collection, was no less data science than the methods and tools of today.

What is different is that today’s data are so vast that one cannot analyze them using hand calculations as Mendel and Copernicus did. New tools, like cloud computing and deep learning, are applied to Fourth Paradigm-sized data streaming in from everywhere, and every day new sources and channels are opening, and more and more data are being collected. An example is the recent launch by the US National Aeronautics and Space Administration (NASA) of the Sentinel-6: ‘Dog Kennel’ satellite used to map the Earth’s oceans (Sentinel-6 Mission, Jet Propulsion Laboratory 2020). As a public institution, NASA makes the data freely available to researchers and the public. It is the availability of these types of data and the rapidity with which they can be accessed that are driving the current scientific revolution.

What are potential candidates for a next paradigm shift? Two related issues immediately come to mind: quantum computing and augmented human cognition. Both require significant

engineering advances and are theoretically possible. Quantum computing (QC) is a revolution occurring in computing (The National Academies of Sciences, Engineering, and Medicine 2019). Like its predecessors supercomputers and cloud computing, QC has the potential to have a huge impact on data science because it will enable scientists to solve problems and build models that are not feasible with conventional computing today. Though QC is not a paradigm shift as such, it would be a tool, like supercomputing before it, that could enable future paradigm shifts and give rise to new ways by which scientists conduct science.

Like earlier paradigm shifts, binary-based, conventional computing (CC) will have a place in big data analysis, particularly regarding interfaces to increasingly powerful and portable sensors. Many scientific analyses do not need the excessive computational capacity that qubit-based QC can provide, and it also is likely that CC will always be more cost-effective to use than QC. The logistical problem of reducing the results of a QC process to one that can be conventionally accessed, stored, and used is just one of the many barriers of using QC today.

Qubits, as opposed to the 1’s and 0’s of binary, can store and compute over exponentially larger problem spaces than binary systems (The National Academies of Sciences, Engineering, and Medicine 2019), enabling the building of models that are challenging to perform today. For example, the ability to compute all potential biological nitrogen fixation in nitrogenase is especially important in that Earth is facing food security problem as the population increases and arable land decreases (Reiher et al. 2017). Nitrogen fixation is a trivial problem for QC analysis that would enable the creation of custom fertilizers for specific food crops that are more efficient and less toxic to the environment. Solving this problem could help end hunger, one of the United Nations Sustainable Development Goals (United Nations 2015).

Another example of a problem that could be addressed by QC would be to monitor and flag hate speech in all public postings across social media sites and enable states to apply legal action

to those who commit such acts within their borders. This problem is beyond CC today and will likely remain so. Yet, many organizations, including the United Nations (United Nations 2020) and the Amnesty International (Amnesty International 2012), have missions to protect humanity against hate speech and its ramifications. This remains a challenge even for the social media companies themselves, who initially rely on participants to flag offensive content and then train systems to identify additional cases, but this method is fraught with challenges (Crawford and Gillespie 2014), the sovereignty, privacy, volume, and changing nature of these data being only a few.

Augmented human cognition (AHC) or human cognitive enhancement (HCE), in addition to QC, is another potential technological capability that will alter the conduct of science. HCE refers to the development of human capabilities to acquire and generate knowledge and understanding through technological enhancement (Cinel et al. 2019). An example of this would be a scientist that no longer needs an external device to perceive light in spectrums that are currently not visible to the human eye, such as ultraviolet. Such a capability would necessitate a scientist to undergo various augmentations, not limited to performance-enhancing drugs, prosthetics, medical implants, and direct human-computer linkages.

HCE is possible because the human brain is estimated to have more processing power than today's fastest supercomputers (Kurzweil 1999) and also will likely have more than near-term QC, although there are researchers investigating the quantum processing "speed limit" (Jordan 2017). The bigger challenge is the neurology – understanding how humans function to a well enough degree to allow the engineering of working human-computer interfaces (Reeves et al. 2007). Overcoming such challenges, assuming they are not legally or politically prevented, could result in significant scientific changes and development.

There are likely to be several other candidates beyond HCE and QC with potential to create a scientific paradigm shift. What these two examples illustrate by looking beyond current abilities to leverage scientific discovery today is that it is not known what the next paradigm shift in science

will be or in what ways it will shift. Revolutionary changes are only conceivable as they emerge or in retrospect.

Further Reading

- Amnesty International. (2012). Written contribution to the thematic discussion on racist hate speech and freedom of opinion and expression organized by the United Nations Committee on elimination of racial discrimination, August 28, 2012. <https://www.amnesty.org/download/Documents/24000/ior420022012en.pdf>. Accessed 9 Dec 2020.
- Anderson, C. (2007) The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 16:07. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory. Accessed 15 Dec 2020.
- Cinel, C., Valeriani, D., & Poli, R. (2019). Neurotechnologies for human cognitive augmentation: Current state of the art and future prospects. *Frontiers in Human Neuroscience*, 13. <https://doi.org/10.3389/fnhum.2019.00013>.
- Crawford, K., & Gillespie, T. (2014). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18. <https://doi.org/10.1177/1461444814543163>.
- Hey, A., Tansley, D., & Tolle, K. (2009). *The fourth paradigm: Data driven scientific discovery*. Redmond: Microsoft Research.
- Jordan, S. P. (2017). Fast quantum computation at arbitrarily low energy. *Physical Review A*, 95, 032305. <https://doi.org/10.1103/PhysRevA.95.032305>.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. New York: Viking.
- NASA. (2020). Juno spacecraft and instruments, NASA website. https://www.nasa.gov/mission_pages/juno/spacecraft. Accessed 8 Dec 2020.
- National Academies of Sciences, Engineering, and Medicine. (2018). *Envisioning the data science discipline: The undergraduate perspective*. Washington, DC: The National Academies Press.
- Q&A with Paul Reville. (2020). The Pandemic's impact on education. *Harvard Gazette*. <https://news.harvard.edu/gazette/story/2020/04/the-pandemics-impact-on-education/>. Accessed 15 Dec 2020.
- Reeves, L. M., Schmorow, D. D., & Stanney, K. M. (2007). *Augmented cognition and cognitive state assessment technology – Near-term, mid-term, and long-term research objectives* (Lecture notes in computer science). In D. D. Schmorow (Ed.) (pp. 220–228). Berlin: Springer.
- Reiher, M., Wiebe, N., Svore, K., Wecker, D., & Troyer, M. (2017). Reaction mechanisms on quantum computers. *Proceedings of the National Academy of Sciences*, 114(29), 7555–7560.

- Sentinel-6 Mission, Jet Propulsion Laboratory. (2020). Sentinel-6: 'Dog kennel' satellite blasts off on ocean mission – BBC News, JPL Website. <https://www.jpl.nasa.gov/missions/sentinel-6/>. Accessed 15 Dec 2020.
- The National Academies of Sciences, Engineering, and Medicine. (2019). *Quantum computing: Progress and prospects*. Washington, DC: National Academies Press.
- United Nations. (2015). The 17 goals | sustainable development. United Nations, Department of Economic and Social Affairs. <http://sdgs.un.org/goals>. Accessed 9 Dec 2020.
- United Nations. (2020). United Nations strategy and plan of action on hate speech. United Nations, Office on Genocide Prevention and the Responsibility to Protect. <https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml>. Accessed 9 Dec 2020.

France

Sorin Nastasia and Diana Nastasia
Department of Applied Communication Studies,
Southern Illinois University Edwardsville,
Edwardsville, IL, USA

Introduction

In recent years, big data has become increasingly important in France, with impact on areas including government operations, economic growth, political campaigning, and legal procedures, as well as on fields such as agriculture, industry, commerce, health, education, and culture. Yet, the broad integration of big data into the country's life has not remained without criticism from those who worry about individual privacy protections and data manipulation practices.

The State of Big Data in France

When François Hollande became the president in 2012, he asserted that big data was a key element of the national strategy aimed at fostering innovation and increasing the competitiveness of the country in the European and the global contexts.

The enthusiasm for big data in the French government resulted in the establishment of the Ministry for Digital Affairs in 2014, under the leadership of notable French socialist party figure and French tech movement member Axelle Lemaire, who reported directly to the Minister of the Economy and Industry at the time, Emmanuel Macron. The Ministry for Digital Affairs was instrumental in devising the Digital Republic Act, a piece of legislation adopted by the National Assembly in 2016, following a widely publicized consultation process. The law introduced provisions to regulate the digital economy, including in regards to open data, accessibility, and protections.

The Digital Republic Act had the goal of providing general guidelines for a big data policy to serve as the basis for further sectorial policies. This landmark piece of legislation made France the world's first nation to mandate local and central government to automatically publish documents and public data. According to this law, all data considered of public interest, including information derived from public agencies and private enterprises receiving support from public funds, should be accessible to the citizenry for free. In a 2015 interview, Axelle Lemaire highlighted the importance of open data as a mine of information for creative ideas to be generated by startups as well as established organizations. She stated: "Both innovation and economic growth are strongly driven by open data from either public or private sources. Open data facilitate the democratic process and injects vitality into society" (Goin and Nguyen 2015). The law also imposed administrative fines of up to 4% of an organization's total worldwide annual turnover for data protection violations.

The legislation was only one part of the project of streamlining the country's core data infrastructure. One component of the data infrastructure is data.gouv.fr, the government portal hosting over 40,000 public datasets from nearly 2000 different organizations. Another component, launched in 2017, is SIRENE, an open register listing legal and economic information about all the companies in France. In 2017, the government of France also launched the Health Data Hub to promote open data and data sharing for supporting medical

and health-care functions, including clinical decision-making, disease surveillance, and population health management. As part of this project, technology specialists are recruited into government, initially on short-term contracts on agile projects followed by full-time employment when the projects mature, at salaries that rival those they could receive in the private sector. However, critics contend that the digitalization of administrative documents and procedures remains unequal among different ministries, and some are largely falling behind (Babinet 2020).

The change of the presidency of France to Emmanuel Macron in 2017 did not diminish the critical significance of big data for the national strategy. In December 2017, the Minister for Europe and Foreign Affairs Jean-Yves le Drian unveiled France's international digital strategy, aimed at serving as a framework for big data use as well as a roadmap for big data practices in regards to government, the economy, and security in international settings. The document highlighted the country's commitment to advocating the inclusion in the digital sphere of states, private sector, and civil societies, spreading digital commons for software, content, and open data, evaluating the impact of algorithms and encouraging their transparency and fairness, educating citizens into the encryption of communications and its implications, fighting the creation and spreading of misinformation, and ensuring the full effect of international law in the cyberspace. The document also expressed support for the concepts of privacy by design and security by design in the way tech products are conceived and disseminated.

Moreover, pundits have claimed that big data helped secure the astounding victory of Emmanuel Macron in his presidential bid in 2017, followed by the triumph of his movement turned into a new political party *La République en marche!* in the French parliamentary elections and local elections held the same year. The campaigns of Macron and *La République en marche!* were supported by data-driven canvassing techniques implemented by LMP, an organization established by Harvard graduates Guillaume Liegey and Vincent Pons and MIT graduate

Arthur Muller who met while volunteering for the Obama campaign in 2008. The three founders of LMP sought to tap into advanced data analytics to reinvent the old-fashioned technique of door-to-door canvassing, devising an algorithm helping French politicians to connect directly to voters. The software package they created, *Fifty Plus One*, flags up the specific characteristics of political territories that need to be targeted. "We were all fascinated by how political parties use data to create mindsets," Liegey stated in an interview (Halliburton 2017).

As France has been developing and testing various approaches based on big data to domestic and international issues as well as to social and political issues, a more skeptical view in regards to analytics aspects of big data has emerged too. In 2016, when France's government announced the creation of a new database to collect and store personal information on nearly everyone living in the country and holding a French identity card or passport, there was immediate outrage in the media. The controversial database, secure electronic documents, was aimed at cracking down on identity theft, but the government's selection of a centralized architecture to store everyone's biometric details raised huge concerns in regards to both the security of the data and the possibilities for misuse of the information. Despite the outcry and the concerns expressed publically, the database was launched and is still in function.

Another area of concern has been judicial analytics. The new article 33 of the Justice Report Act adopted in 2019 makes it illegal to engage in the use of statistics and machine learning to understand or predict judicial behavior. The law states: "No personally identifiable data concerning judges or court clerks may be subject to any reuse with the purpose or result of evaluating, analyzing, or predicting their actual or supposed professional practices." The article applies to individuals as well as technology companies and establishes sanctions for prejudices pertaining to data processing. The law is supported by civic society groups and legislators believing that the increase in public access to data, even with some privacy protections in place, may result in unfair and discriminatory data manipulation and

interpretation. However, it has also had its opponents. “This new law is a complete shame for our democracy,” stated Louis Larret-Chahine, the general manager and cofounder of Prédicite, a French legal analytics company (Livermore and Rockmore 2019).

Both governmental and nongovernmental entities in France have demanded the monitoring and, when needed, action against companies using consumer data for the personalization of content, the behavioral analysis of users, and the targeting of ads. In 2019, France’s data protection authority established through the Digital Republic Act fined Google 50 million Euros for the intrusive ad personalization systems and its inadequate systems of notice and consent when users create accounts for Google services on Android devices. The data protection authority found that Google violated its duties by obfuscating essential information about data processing purposes, data storage periods, and categories of personal information used for ad personalization. The decision implied that behavioral analysis of user data for the personalization of advertising is not necessary to deliver mail or video hosting services to users.

Conclusion

In France, big data policy was considered by the Hollande administration and has continued to be considered by the Macron administration as a potential direct economic growth driver and an

opportunity to establish the nation as a pioneer in regards to digital processes in global settings. France has become broadly thought of as one of the best in the world for open data, while the European data portal proclaimed it one of the trendsetters in European data policy. While the government of France has succeeded to build a strong belief that data openness and data processing for the benefit of people is key to leading government and business digital transformation, concerns remain in regards to privacy as well as to uses of data by such organizations as courts and businesses.

Further Reading

- Babinet, G. (2020). The French government on digital – Midterm evaluation. *Institut Montaigne*. <https://www.institutmontaigne.org/en/blog/french-government-digital-mid-term-evaluation>
- Goin, M., & Nguyen, L. T. (2015). A big bang in the French big data policy. <https://globalstatement2015.wordpress.com/2015/10/30/a-big-bang-in-the-french-big-data-policy/>
- Halliburton, R. (2017). How big data helped secure Emmanuel Macron’s astounding victory. *Prospect Magazine*. <https://www.prospectmagazine.co.uk/politics/the-data-team-behind-macrons-astounding-victory>
- Livermore, M., & Rockmore, D. (2019). France kicks data scientists out of its courts. *Slate*. <https://slate.com/technology/2019/06/france-has-banned-judicial-analytics-to-analyze-the-courts.html>
- Toner, A. (2019). French data protection authority takes on Google. *Electronic Frontier Foundation*. <https://www.eff.org/deeplinks/2019/02/french-data-protection-authority-takes-google>

Gender and Sexuality

Kim Lorber¹ and Adele Weiner²

¹Social Work Convening Group, Ramapo College of New Jersey, Mahwah, NJ, USA

²Audrey Cohen School For Human Services and Education, Metropolitan College of New York, New York, NY, USA

Introduction

Gender refers to ways one self-defines as male and female or within this spectrum based on social constructs separate from biological characteristics. Sexuality is one's expression of their sexual identity. Individuals can be heterosexual, homosexual, or another along a fluid spectrum that can change throughout one's life. In some societies one's gender or sexuality offers different social, economic, and political opportunities or biases.

Big data is a useful tool in understanding individuals in society and their needs. However, the multitude of available resources can provide a transparency to one's private life based on technology. Regardless of whether the information is gathered by retailers or the federal government, profiles can include much private information ranging from television preferences to shopping profiles all of which combined can include very personal information related to one's gender identity and sexuality.

A variety of resources, including social network profiles and shared photos, can provide group affiliations, favorite films, celebrities, friendships, causes, and other personal information. Facebook privacy concerns arise regularly. While intended for friends to see, data miners can easily draw conclusions about an individual's gender and sexuality. Short of living off the grid without traceable billing, banking, and finances, some residue of much we do in our lives is collected, somewhere. This entry proposes to demonstrate how current and future research will be based on very private elements of our lives in regard to attitudes toward and experiences of gender and sexuality.

Big Data on Gender and Sexuality

The immediacy and interconnectedness of big data become obvious when one finds an item looked at on Amazon.com, for example, as a regular advertisement on almost every accessed webpage with advertising. Big data calculations create recommendations based on likelihoods about one's personal lifestyle, gender, sexual orientation, entertainment, and retail habits. Emails abound with calculated recommendations. This can easily feel like an invasion of privacy. Is there a way to exclude oneself and how can such personal information challenge one's life? Who has access to it?

Gender and sexuality are very personal to each individual. Most people do not blatantly discuss such private elements of their lives. Ideally, a utopian society would allow anyone to be whoever they are without judgment; our society is not this way, and implications from disclosure of certain types of information can result in bias at social, professional, medical, and other levels. Such biases may not be legal or relevant and can lead to detrimental results.

Friends, strangers, or casual acquaintances may know more about one's personal gender and sexuality details than spouses, other family members, and best friends. Can personal information be disclosed by personalized web advertising viewed by another while one innocently concentrates on elements of interest not even noticing the other uninvited clues about their life?

The Internet had been a forum for individuals to *anonymously* seek information and connections, explore friendships and relationships, and live and imagine beyond their developed social identities. This has been particularly true for individuals seeking information about gender and sexuality, in a society where gender identity is presumed as fixed and discussions of sex may be taboo. Information now readily extrapolated by big data crunchers can make an environment unsafe and no longer anonymous open for anyone to find many of one's most personal thoughts and feelings. For example, how much information can be found about someone's gender identity or sexual orientation via access to their online groups joined or "liked" on Facebook, followed on Twitter, etc. and, combined perhaps with retail purchases, consultations, or memberships, and social venues, which can strongly suggest another layer to an individual's calculated and assumed private identity?

Similarly, virtual friendships and relationships, groups, and information sites can be calculated into gender and sexual identity conclusions. If a man married to a woman purchases books, borrows similar titles from a library, joins an online group, and makes virtual friends who may be questioning their heterosexuality, a trail has been created. Perhaps uncomfortable exploring these concerns privately, advertisements on websites

visited may now offer other resources which can be jarring, and there is no easy way to make them disappear.

Big Data and Gender

Big data analysis can also provide answers to long overlooked biases and inequities. Melinda Gates (2013) discussed eight Millennium Development Goals (MDGs), adopted by the UN in 2000 to serve as a charter for the field of development. Gender is not one of the big data items specifically explored beyond eliminating gender differences in education. Ideally new priorities post-2015 will address gender-based areas such as violence, property rights for women, and the issue of child marriage. Increasingly precise goals addressing women and agriculture, for example, would show globally the strengths and weaknesses by country presenting big data conclusions. In sub-Saharan Africa, women do the majority of farm work. However, agricultural programs are designed for a minority of male farmers possibly because government trainers either prefer working with men or are not allowed to train women resulting in female farmers being significantly less productive. Big data highlighting the disparity between the genders in such programs will be used to make programs more equitable.

Big Data and Sexual Identity

Google, a seemingly limitless and global information resource, allows for endless searching. Using the search autofill feature, some tests were done using key words. The top autofill responses to *homosexuality should* resulted in: "homosexuality should be banned," "homosexuality should be illegal," "homosexuality should be accepted," and "homosexuality should be punishable by death." "Gay men can't" "*give blood*" was a top autofill suggestion. "Bisexuality is" was autofilled with "bisexuality is not real," "bisexuality is real," and "bisexuality isn't real." Many of these autofill responses, based on past searches, reflect biases

and stereotypes related to gender and sexual orientation.

Kosinski et al. (2013) studied Facebook *likes* and demographic profiles of 58,000 volunteers. Using logistic/linear regressions, the model was able to identify homosexuals versus heterosexuals (88% accuracy).

Conclusion

While great outcomes are possible from big data, as in the case of identifying discrimination based on gender or sexual orientation, there is also the risk of culling deeply personal information, which can easily violate gender and sexuality privacy. Like any other personal information, once it is made public, it cannot be taken back. Big data allows the curious perhaps absent negative intentions, to know what we, as citizens of the world, wish to share on our own terms. The implications are dramatic and, in the case of gender and sexual identity, can transform individuals' lives in positive or, more likely, negative ways.

Cross-References

- [Data Mining](#)
- [Data Profiling](#)
- [Privacy](#)
- [Profiling](#)

Further Reading

- Foremski, T. (2013). Facebook 'Likes' can reveal your sexuality, ethnicity, politics, and your parent's divorce. Retrieved on 28 Aug 2014 from www.zdnet.com/facebook-likes-can-reveal-your-sexuality-ethnicity-politics-and-your-parents-divorce-7000013295/.
- Gates, M. (2013). *Bridging the gender gap*. Retrieved on 28 Aug 2014 from http://www.foreignpolicy.com/articles/2013/07/17/bridging_the_gender_gap_women_data_development.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 110(15), 5802–5805.

Genealogy

Marcienne Martin

Laboratoire ORACLE [Observatoire Réunionnais des Arts, des Civilisations et des Littératures dans leur Environnement] Université de la Réunion, Saint-Denis, France, Montpellier, France

Articulated around the “same” and the “different,” the living world participates in strange phenomena: architect genes, stem cells, and varied species whose specificity is, however, derived from common trunks. Each unit forming the said universe is declined, simultaneously, into a similar object and a completely different object (Martin 2017). Through different genetic researches, geneticists have created new technologies as CRISPR (<https://www.aati-us.com/applications/crispr-cas9/>) (*Clustered Regularly Interspaced Short Palindromic Repeats*) which allow the decoding of the transformation of a stem cell into a specific cell and the evolution of this last. These new technologies correspond to a database and, indirectly, to big data.

The lexical-semantic analysis of the term “genealogy” refers, firstly, to the study of the transmission chains of genes; secondly, this lexical unit also refers to anthroponomy, which is the mode of symbolic transmission of the genetic structure through nomination in human societies. Moreover, beyond the symbolic transmission of genetic codes received through nomination, civil society is based on the transmission of property to descendants or to collaterals. Finally, genealogy also points to different beliefs in relation to the world of death.

The basis of genetics was founded by the discoveries made, in particular, by Johann Gregor Mendel (1822–1884) who studied the transmission of genetic characteristics with respect to plants (peas). These results formed the basis of what is now known as Mendel's laws which define how genes are transmitted from generation to generation. Furthermore, Darwin showed in his analysis of the evolution of species in the living world, how a variety of life may be part of a

species, based on the differences and similarities as well as the adaptation to a specific environment. These scientific approaches have opened new fields of research in relation to the human world. In his entomological study, Jaisson (1993) states that in relation to bees some genetic markers are used to supply different behavioral predispositions among workers, such as reflex conditioning, conditional or acquired, according to Ivan Petrovich Pavlov. According to Dawkins (1990), who is a neo-Darwinian ethologist, evolution occurs step by step, for the survival of certain genes and the elimination of others in the gene pool. These studies raise the question of whether certain phenomena in human beings are innate and acquired.

Anthroponomy is the identity marker for the transmission of such genetic codes in such a human group. Another function of anthropogenesis is to find a name for a new object in the world in analogy with an existing object integrated in the human paradigm. As Brunet et al. (2001–2002) pointed out, every individual is carrying a name that refers to his or her community, but also which is a reference to his or her cultural practices; the name will not be the same and will not be transmitted in the same mode, depending on the different human groups. Furthermore, Martin (2009) states that “the other” considered in its only otherness can also integrate divine paradigms. Louis XIV, who was not only named the Sun King, had beside his status of a monarch also divine right. In relation to anthroponomy, identity is linked to one’s descent. Indeed, if one has to name an object in the world, one also has to give it a sense and to identify an individual means to recognize him or her, but also to put him or her into the group they belong to. The first group an individual belongs to is that of his or her gender, which is administratively materialized when registered as one’s civil status after one’s birth. In French society, in addition to gender anthroponyms (full name, surname), date and place of birth as well as the identity of the parents are registered (Martin 2012).

From one culture to another, anthropogenesis may take various turns, sometimes connected to practical considerations, such as in groups that do

not name the newborn until he or she reaches an age where his or her survival is most likely. In relation to the nomination of newborn children, Mead (1963), an anthropologist who studied different ethnic groups, including Arapesh, specifies that among the Arapesh, as soon as a newborn smile when looking at his or her father, he or she will be given a name, in particular a name of a member of the father’s clan. The creation that presides over the implementation of the *nomen* and of its transmission is often articulated around the genealogical chain. Thus, there are systems called “rope,” which are either links which connect a man, his daughter, and his daughter’s sons or a woman, her son, and his son’s daughters (Mead, 1963). Ghasarian (1996), in a study on kinship, mentions that the name is not automatically given at birth. So, Emperaire (1955), a researcher at the Musée de l’Homme in Paris, gives the example of Alakalufs, an ethnicity living in Terre de Feu, which does not give any names to newborns; indeed, at their birth children do not receive a name; it is only when they begin to talk and walk that the father chooses one.

Anthroponomy may also point to one’s family history. For example, Levi-Strauss (1962) gives the example of the Penans, a tribe of nomads living in Borneo, where at the birth of their first child, the father and the mother adopt a “tecknonym” expressing their relationship to their child specifically designated such as *Tama Awing*, *Tinen Awing*: father’s (or mother’s) Awing. Levi-Strauss also states that Western Penans have no less than 26 different necronyms corresponding to the degree of relationship, according to the deceased’s age, the gender, and also to the birth order of children until the ninth.

To write the story of the life, tale of a newborn begins at his or her identification; so in the Vietnamese culture, the first emblematic name given to the Vietnamese child is used for his or her private family use. This is not always a beautiful name, but a substantivized adjective that emerges according to the event or to the family experience when the baby is born. There was also a denomination created to disabuse an evil genius who could make the child more fragile when hearing

his or her beautiful name. The name given to the child usually has a meaning, a moral quality, or it is the name of an element of nature whose literature makes it a symbol. In Vietnam which is a tropical country, the name *tuyet* means snow and it is given in reference to its whiteness as a symbol of purity and of sharpness as the literature inspired by the Chinese culture reports (see Luong, 2003). In the Judeo-Christian culture, hagiography refers to the exemplary life of such a person considered in the context of this religious ideology. Furthermore, Héritier (2008), referring to the different kinship systems, specifies that it simply gives us indications on the fact that all human populations have thought about the same biological data, but they have not thought about it the same way. Thus, in patriarchal societies, the group, or the individual, has the status of the dominant male, while in matriarchal societies it is women who are the ones that integrate a dominant role.

Anthroponomy in the context of genealogy is a procedure that contributes to the creation of the identity of the individual. All of these modes allow the social subject to distance themselves from the group and to become an individual whose status is more or less affirmed according to a more or less meaning full group structure. And it is based on these procedures that identity takes its whole significance. The nominal identity incorporates de facto the individual in his or her genealogical membership group, irrespective of whether the latter is real or based on adoption. The process, inferring the identity structure of a social subject, seems to be initiated by the nomination act. This identity also reflects the place occupied by the family in the social group. The sets and the alliances of family are the subject of limited arrangements as analyzed by Héritier (2008). The reason is that the rules we follow locally to find a partner are adapted to systems kinship groups classifications, as to kinship and alliance; they can be subtly altered by the games of power, witchcraft, and economy. Diamond (1992) showed that we share over 98% of our genetic program with primates (pygmy chimpanzee of Zaire and the common chimpanzee from Africa). This has as a consequence

in the appearance of certain types of similar behaviors, as reflected by the establishment of dominant groups with alpha males, like that of nobility, especially royalty.

Nobility refers to a highly codified patronymic system referring to the notion of belonging. According to Brunet and Bideau when identifying (2001–2002) particular groups within society, they generally include several elements: a surname (often prior to the ennoblement) and one or more land names and titles (in *Le patronyme, histoire, anthropologie et société*, 2001–2002). If we refer to the book by Debax (2003), who in her study focuses on the feudal system in the eleventh and the twelfth centuries in the Languedoc (region of France), it is shown that the relationship between lords and vassals is articulated around the concept of fiefdom whose holding always involves at least two individuals, one holding the other, which results in the importance of the personal relationship when the specification is a fiefdom anthroponym. Hence, the term “fief” was used by Bernard d’Anduze, Bermond Pelet, and Guilhem de Montpellier. The particle “de” is a preposition, which in this case marks the origin as “calissons,” sweets made in Aix-en-Provence (France). As part of this type of anthropogenesis, we have several layers of onomastic formation whose interpenetration emphasizes different levels of hierarchical organization. In France, these names are often composed of a first name called by interested “surname” and a second, or even a third name, called “land names,” connected by a preposition (e.g., *Bonnin de la Bonnière de Beaumont*); these names usually place their holders in a category of nobility, who has been legally abolished for more than two centuries; nevertheless it has remained important in the French society (Barthelemy in *Le patronyme, histoire, anthropologie et société*, 2001–2002).

The death is part of an underlying theme in life through the use of some anthroponomical forms as the necronyms cited by Lévi-Strauss (1962); this mode of nomination expresses the family relationship of a deceased relative to the subject. The *postmortem* identification of the social subject is associated the memorial inscription that is found, in particular, on the plates of memorials

and of gravestones. This discursive modality seeking to maintain, in a symbolic form, the past reality of the individual, also finds its realization in expressions like Mrs. X (Y widow) wife . . . late Mr. X., the remains of Mr. Z, or it can take reality in the name given to a newborn which has been given by a member of a group of deceased. The cult of ancestors is similar to the ceremonial practices addressed collectively to the ancestors belonging to a same lineage, often done on altars; this is another way to keep ascendants of the genealogical relationship in the memory of their descendants. Finally, the construction of a genealogical tree, which is very popular in Western societies, refers to a cult of symbolic ancestors.

Further Reading

- Brunet, G., Darlu, P., & Zei, G. (2001–2002). *Le patronyme, histoire, anthropologie et société*. Paris: CNRS.
- Darwin, C. (1973). *L'origine des espèces*. Verviers: Marabout Université.
- Dawkins, R. (1990). *Le gène égoïste*. Paris: Odile Jacob.
- Debax, H. (2003). *La féodalité languedocienne XIe – XIIe siècle – Serments, hommages et fiefs dans le Languedoc de Trencavel*. Toulouse-Le Mirail: Presses Universitaires du Mirail.
- Diamond, J. (1992). *Le troisième singe – Essai sur l'évolution et l'avenir de l'animal humain*. Paris: Gallimard.
- Empeiraire, J. (1955). *Les nomades de la mer*. Paris: Gallimard.
- Ghasarian, C. (1996). *Introduction à l'étude de la parenté*. Paris: Éditions du Seuil.
- Héritier, F. (2008). *L'identique et le différent*. Paris: Diffusion Seuil.
- Jaisson, P. (1993). *La fourmi et le sociobiologiste*. Paris: Odile Jacob.
- Lévi-Strauss, C. (1962). *La pensée sauvage*. Paris: Plon.
- Luong, C. L. (2003). L'accueil du nourrisson: la modernité de quelques rites vietnamiens. *L'Information Psychiatrique*, 79, 659–662. <http://www.jle.com/fr/revues/medecine/ipe/e-docs/00/03/FC/08/article.md>.
- Martin, M. (2009). *Des humains quasi objets et des objets quasi humains*. Paris: Éditions L'Harmattan.
- Martin, M. (2012). *Se nommer pour exister – L'exemple du pseudonyme sur Internet*. Paris: Éditions L'Harmattan.
- Martin, M. (2017). *The pariah in contemporary society: A black sheep or a prodigal child?* Newcastle upon Tyne: Cambridge Scholars Publishing.
- Mead, M. (1963). *Mœurs et sexualité en Océanie – Sex and temperament in three primitive societies*. Paris: Plon.

Geographic Information

► Spatial Data

Geography

Jennifer Ferreira
Centre for Business in Society, Coventry
University, Coventry, UK

Geography as a discipline is concerned with developing a greater understanding of processes that take place across the planet. While many geographers agree that big data presents opportunities to glean insights into our social and spatial world, and the processes that take place within it, many are also cautious about how it used and the impact it may have on how these worlds are analyzed and understood. Given that often big data are either explicitly or implicitly spatially or temporally referenced, this makes it particularly interesting for geographers. Geography, then, becomes part of the big data phenomenon.

As a term that has only relatively recently become commonly used, definitions of big data still vary. Rob Kitchin suggests there are in fact seven characteristics of big data, extending beyond the three Vs proffered by Doug Laney (volume, velocity, and variety) which are widely cited:

1. Volume: often terabytes and sometimes petabytes of information are being produced.
2. Velocity: often a continuous flow created in near real time.
3. Variety: composed of both structured and unstructured forms.
4. Exhaustivity: striving to capture entire populations.
5. Fine grained: aiming to provide detail.
6. Relational: with common fields so data sets can be conjoined.

7. Flexible: so new fields can be added as required, the data can be extended and where necessary exhibit scalability.

This data is produced largely through three forms: directed, generated largely by digital forms of surveillance; automated, generated by inherent automatic functions of digital devices; and volunteered, provided by users, for example, via interactions on social media or crowdsourcing activities.

The prevalence of spatial data has grown massively in recent years, with the advent of real-time remote sensing and radar imagery, crowdsourcing map platforms such as OpenStreetMap, and digital trails created by ubiquitous mobile devices. This has meant there is a wealth of data to be analyzed about human behavior, in ways not previously possible.

Large data sets are not a new concept for geography. However, even some of the most widely used large data sets used geography, such as the census, do not constitute big data. While they are large in volume and seek to be exhaustive, and high in resolution, they are very slow to be generated and have little flexibility. The type of big data now being produced is well exemplified by companies such as Facebook which in 2012 alone processed over 2.5 billion pieces of content, 2.7 billion “likes,” and 300 million photo uploads in just 1 day or Walmart which generated over 2.5 petabytes of data information every hour in 20,102. One of the key issues for using big data is that collecting, storing, and analyzing these kinds of data is very different from that of traditionally large data sets such as the census. These new forms of data creation are creating new questions about how the world operates, but also about how we analyze and use such data forms.

Governments are increasingly turning to big data sources to consider a variety of issues, for example, public transport. A frequent system referred to about the production of big data related to public systems is the use of the Oyster card in London. Michael Batty discusses the example of public in transport in London (tube, heavy rail, and buses) to consider some of the issues with big

data sets. With around 45 million journeys every week or around a billion every year, the data is endless. He acknowledges that big data is enriching our knowledge of how cities function, particularly with respect to how people move around them. However, it can be questioned how much this data can actually tell us. Around 85% of all travelers using public transport in London on these forms of transport use the Oyster card, and so clearly there is an issue about representativeness of the data. Those that do not use the card, tourists, occasional users, and other specialist groups will not be represented. Furthermore because we cannot actually trace where an individual begins and ends their journey, it only presents a partial view of the travel geographies of those in London. Nevertheless this data set is hugely important for the operation of transport systems in the city.

Disaster response using big data has also received significant media attention in recent years: crisis mapping community after the 2010 Haiti earthquake or collecting tweets in response to disaster events such as Hurricane Sandy. This has led to many governments and NGOs promoting the use of social media as potentially useful data sources for responding to disasters. While geo-referenced social media provides one lens on the impact of disaster events, it should not be relied on as a representative form of data covering all populations involved. Big data in these scenarios presents a particular view of the world based on the data creators and essentially can mask the complexity and multiplicity of scenarios that actually exist. Taylor Shelton, Ate Poorthuis, Mark Graham, and Matthew Zook explore the use of twitter around the time of Hurricane Sandy, and they acknowledge that their research did not present any new insights into the geography of twitter, but that it did show how subsets of big data could be used for particular forms of spatial analysis.

Trevor Barnes argues that criticisms of the quantitative revolution in geography are also applicable to the use of big data. First that a focus on the computational techniques and data collected can become disconnected from what is

important, i.e., the social phenomena being researched. Second that it may create an environment where quantitative information is deemed superior and that where phenomena cannot be counted they will not be included. Third, that numbers do not speak for themselves – numbers created in data sets (of any size) emerge as a product of particular social constructions even where they are automatically collected by technological devices.

The growth of big data as part of the data revolution presents a number of challenges for geographers, there has been much hype and speculation over the adoption of big data into societies, changing the ways that businesses operate, the way that government manage places, and the way that organizations manage their operations. For some, the benefits are overstated. While it may be assumed that because much technology contains GPS that the use of big data sets is a natural extension of the work geographic information scientists, it should be noted that the emergence of such data sets created by mobile technology has created a large new amount of data, but also data that geographic information scientists have not typically focused their efforts. Therefore work is needed to develop sound methodological frameworks to work with such data sets.

The sheer size of the data sets that are being created, sometimes with millions or billions of observations being created in a variety of forms on a daily basis, is a clear challenge. Traditional statistical analysis methods used in geography are designed to work using smaller data sets with much more known about the properties of the data being analyzed. Therefore new methodological techniques for data handling and analysis are required to be able to extract useful information for geographical research.

Data without a doubt are a key resource for modern the world; however it is important to remember that data does not exist independently of the systems (and people in them) from which they are produced. Big data sets have their own geographies; they are themselves social

constructions formed from variegated socioeconomic contexts and therefore will present a vision of the world that is uneven in its representation of populations and their behavior. Big data, despite attempts to make it exhaustive, will always be spatially uneven and biased. Data will always be produced by systems that have been created with influences from different contexts and from groups of people with different interests.

Sandra González-Bailón highlights that while technology has allowed geospatial data to be generated much more quickly than in the past, and if mobilized in an efficient manner, people can use these technologies as network of sensors. However, the use of digital tools can produce distorted maps or results if the inputs to the system are systematically biased, i.e., those who do not have access to the tools will not be represented. Therefore there are questions about how to extract valid information from the ever-growing data deluge. Then there are issues around privacy and confidentiality of the data produced and how it will be used potentially in both the public and private sectors.

Michael Goodchild highlights that while a lot of big data is geo-referenced and can contribute to a growing understanding of particular locations, there are issues about quality of data that is produced. Big data sets are often comprised of disparate data sources which do not always have quality controls or do not have metadata about the provenance of the data. This raises questions about the extent such data can be trusted, or used, to make valid conclusions. There is a need for geographers to explore how data can become more rigorous. Michael explains how Twitter streams continue to be seen as important sources of information about social movements, or events, but often little is known about the demographics of those tweeting, and so it is impossible to understand the extent to which these tweets represent the wider sentiments of society. Furthermore, only a small percentage of tweets are geo-referenced, and so the data is skewed toward the data provided by people who opt in to provide that level of data. Much like many other geographers writing on the topic of

big data, the potential for such sources of data to be useful, but questions need to be raised about how it is used and how the quality is improved.

Mark Graham has begun to ask questions about the geographies of big data and considered which areas of the world are displayed through big data sets and what kinds of uneven geographies are produced by them. The geographies of how data is produced are revealing in itself. This is exemplified by examining the content of Wikipedia: every article on Wikipedia was downloaded and placed on a map of the world. While this showed a global distribution, particularly for articles in English language, the worlds displayed by those in Persian, for example, were much more limited. The key point here was that the representations made available to the world through the use of big data can lead to the omission of other worlds that still exist but may not be visible. These absences or “data shadows” are also a concern for geographers. It raises questions about what this says about the places they represent. In exploring this phenomenon, geographers are seeking to explore the geographies of data authorship in the data deluge, considering why there are differences in the production of information and asking questions about why some people produce large amounts of data while others are excluded.

It is without question that digital technologies have transformed the ways in which we can explore the way the world works; the flood of data now being produced can be used to create more maps of places, more models of behavior, and more views on the world. With companies, governments, and research funding agencies calling for more effort to be put into generating and exploring big data, some geographers have highlighted that in order to deliver significantly valuable insights into societal behavior, then more effort is needed to ensure that the big data collection and analysis are scientifically robust. Big data and particularly data that are geo-referenced have provided a new wealth of opportunities to understand more about people and places, asking new questions and measuring new processes and phenomena in ways not previously possible.

Cross-References

- [Demographic Data](#)
- [Disaster Planning](#)
- [Smart Cities](#)
- [Socio-spatial Analytics](#)
- [Spatial Data](#)

Further Reading

- Barnes, T. (2013). Big data, little history. *Dialogues in Human Geography*, 3(3), 297–302.
- Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3), 274–278.
- Gonzalez-Bailon, S. (2013). Big data and the fabric of human geography. *Dialogues in Human Geography*, 3(3), 292–296.
- Goodchild, M. (2013). The quality of big (geo)data. *Dialogues in Human Geography*, 3(3), 280–284.
- Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), 262–267.
- Kitchin, R. (2014). *The data revolution: Big data open data, data infrastructures and their consequences*. London: Sage.
- Li, L., Goodchild, M., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of twitter and Flickr. *Cartography and Geographic Information Science*, 40(2), 61–77.
- Laney, D. (2001). 3D data management: controlling data volume, velocity, and variety. Available from: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3DData-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> Accessed 18 Nov 2014.
- Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of hurricane Sandy: Uncovering the sociospatial dimensions of ‘big data’. *Geoforum*, 52(1), 167–179.

Geospatial Big Data

- [Big Geo-data](#)

Geospatial Data

- [Spatial Data](#)

Geospatial Information

► Spatial Data

Geospatial Scientometrics

► Spatial Scientometrics

Google

Natalia Abuín Vences¹ and Raquel Vinader Segura^{1,2}

¹Complutense University of Madrid, Madrid, Spain

²Rey Juan Carlos University, Fuenlabrada, Madrid, Spain

Google Inc. is an American multinational company specialized in products and services related to the Internet, software, electronic devices, and other technology services, but its main service is the search engine that gives name to the company and according to data from Alexa is the most visited site in the world.

The company was founded by Larry Page and Sergey Brin. These two entrepreneurs met each other while studying at Stanford University in 1995. In 1996, they created a search engine (initially called BackRub) that used links to determine the importance of specific web pages. They decided to call the company Google, making a play on the mathematical term “googol” used to describe a number followed by a hundred zeros. Google Inc. was born in 1998 when Andy Bechtolsheim, cofounder of Sun Microsystems, wrote a check for \$ 100,000 to this organization, which until now did not exist. At present the company has more than 70 offices across 40 countries and has over 50,000 employees.

The corporate culture of the organization is geared toward the care and trust of human resources. In 2013 Google was elected for the

fourth time as the best company to work for in America in a list compiled by Fortune. Their offices were designed to work in the company as in the place where it was germinated: a college campus.

People are what really make the company that Google is. They hire smart and determined people, and they set above the capacity for working to the experience. While Googlers (this is how the employees of this company are known) share goals and expectations about the company, come from diverse professional fields, and among all speak dozens of languages, they represent a global audience for which they work.

Google keeps an open culture that usually occurs at the beginning of a company, when everyone contributes in a practical way and feels comfortable sharing ideas and opinions. Googlers do not hesitate to ask questions on any matter of the company directly to Larry, Sergey, and other executives in both the Friday meetings and email and in the coffee shop.

The offices and coffee shops are designed to promote interaction between Googlers and encourage work and play. The offices offer massage services, fitness center, and a wide range of services that allow workers to relax and interact with each other.

The company receives over one million resumes per year, and the selection process may extend over several months. Only 1 out of 130 applicants gets a place in the company, while at Harvard, one of the best universities in the world gets it out 1 of 14.

Google: The Search Engine

The centerpiece of the company is the search engine, which processes more than two million requests per second, compared to the 10,000 that processed daily during the year of its creation.

The Google search index has a size of more than 100 million gigabytes. To simplify the data, 100,000 1 TB hard drives are needed to reach this capacity. Google aims to improve the daily load time searching and indexing more URLs and improve search.

Each month the search engine receives more than 100 million unique users that perform 12,800 million searches.

PageRank is one of the successful search engines, a family of algorithms used to numerically assign the relevance of web pages indexed by a search engine.

PageRank relies on the democratic nature of the web by using its vast link structure as an indicator of the value of a particular page. Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives, and also analyzes the page that casts the vote. The votes cast by an “important” page, i.e., with high PageRank weight more, help to consider other “important” page. Therefore, the PageRank of a page reflects the importance of the Internet itself. Thus the more links a site receives from others, the more lasting are those links come from reputed spaces, and the greater the chances of a page appear in the top search results of Google.

The route of a search begins long before inserting it into Google. The search engine uses computer robots (spiders or crawlers) looking web pages for its inclusion in the Google search results. Google software stores data on these pages in data centers. The web is like a book with billions of pages that are being indexed by the search engine, a task to which they have already spent more than one million hours.

When the user initiates a search, the Google algorithm starts searching for the information demanded by the Internet. The search runs an average of 2400 miles before offering an answer: it can stop in data centers around the world during the journey and travels hundreds of millions of miles per hour traveling at nearly the speed of light.

When the user enters a query, it would appear predicted searches and results before pressing Enter. This feature saves time and helps to get the required response faster. This is what is called Google Instant.

This feature has several advantages:

- Faster searches: predicted searches show results before the user has finished drafting it; Google Instant can save 2–5 s per search.

- More accurate predictions: even if the user does not know exactly what he is looking for, predictions will guide him in his search. The first prediction is shown in gray so the user can stop typing as soon as he finds what he is looking for.
- Instant results: when user starts writing, the results begin to appear.

The algorithm checks the query and uses over 200 variables to choose the most relevant answers among millions of pages and content. Google updates their ranking algorithms with more than 500 improvements annually.

Examples of the variables used to locate a page are:

- Updating content of a website
- Number of websites that target a home page and links
- The website keywords
- Synonyms of keywords searched
- Spelling
- Quality of the content of the site URL and title of the web page
- Personalization
- Recommendations of users to whom we are connected

The results are sorted by relevance and displayed on the page. In addition to instant results, we can get a preview of web pages by placing the cursor on the arrows to the right of a result to quickly decide whether we want to visit that site.

The Instant Preview takes a split second to load (on average).

Some important figures related to the search are listed:

- Google has responded 450.000 million new queries (searches never seen before), since 2003.
- 16% of daily searches are new.

AdWords

The main source of income of the company and the search engine is advertising. In 2000 Google

introduced AdWords, a method of dynamic advertising for the client: advertisers, with the concept of pay-per-click ads, pay only for those in which a surfer clicked. On the part of the site owners, they charge based on the number of clicks the ads on their website have been generated.

Running AdWords is simple: the advertiser writes an ad showing potential customers the products and services he offers. Then choose the search terms that will make this announcement to appear among Google search results. If the key words entered by users match the search terms selected by the advertiser, the advertising is displayed above or alongside search results. This tool generates 97% of company revenues.

Other Products and Services

As already mentioned above, the search engine is the flagship product of the company, in which spend most of their human and material resources. But the company has a great capacity for innovation and has developed many products and services to market in the field of Information and Communication Technologies related to social media and communication, maps, location, web browsing, developers, etc.

Then we will classify according to the type of functionality, major Google products, and services.

Services and Information Searches

In addition to the web search engine, Google offers a number of specialized services that can locate only certain kinds of content:

- Google Images: searching only images and pictures stored on the web.
- Google News: shows the latest news on any topic, introducing a term. It allows us to subscribe and receive them by email. This way, users will get this news tracking several websites dedicated to information.
- Google Blog Search: localized content hosted on blogs.
- Google Scholar: searches for books published in academic journals and books.

- Google Books: allows to locate books on the web.
- Google Finance: seeks economic news.

Social and Communication Services

- Google Alerts: this is a notification system through email that sends alerts based on a search term when it appears in search engine results.
- Google Docs is a tool for creating and sharing documents, spreadsheets, and presentations online. It works similar to the programs of the office suite.
- Google Calendar is a tool that lets you organize and share online calendar appointments and events, and it is integrated with Gmail. This tool is available to access data from computers or other devices.
- Google Mail (Gmail) is the main and most important email service in the Internet, with all the necessary features. Its speed, safety features, and options make it almost unnecessary to use another email service. In addition, through this mail service, the user has direct access to many of the services and products such as Google+, Google Drive, Calendar, YouTube, etc.
- Google Plus (Google+) is the social network of Google and in the first half of 2014 had nearly 350 million active users. This service is integrated into the main services of Google. When we create a Google account, we automatically are part of this network.
- Google Hangouts: this is an instant messaging service that also allows us to make video calls and conferences between several people. It began as a text chat with video support up to ten participants and to take control of SMS devices with Android.
- Google Groups: this application allows us to create discussion groups on almost any topic. Users can create new groups or join existing ones, make posts, and share files.
- Play Google: the Android store content. We can find millions of apps, books, movies, and songs, paid and free.

Maps, Location and Exploration

- Google Maps provide different maps from around the world and allow us to calculate distances between different geographic locations.
- Google Street View displays panoramic photos of places and sites from GMaps.
- Google Earth: this is a virtual 3D World Atlas, used for satellite imagery and aerial photos. There are applications for portable devices. Other similar versions let the user admire the moon and Mars and explore space using photographs from NASA: Google Moon, Google Mars, and Google Sky.

Tools and Utilities

- Google Translate is a complete service to translate text or web pages in different languages. We can have the offline service with Android devices.
- Google Chrome: the Google web browser. It is simple and fast and has a lot of extensions to add functionality. All our navigation data such as history, saved passwords, bookmarks, and cache can be synchronized automatically with our Google account so we have a backup in the cloud and have them in different computers or devices.
- Android is an open operating system for portable device code.
- Google AdSense is an advertising service to display ads on the pages of a website or a blog created in blogger. The ads shown are from the Google AdWords service.
- Blogger is a publishing platform that lets us create personal blogs with their own domain name.
- Google Drive is a storage service in the cloud to store and share files such as photos, videos, documents, and any other file types.
- Picasa is an application that lets us store photos online, linked to our account in the social network Google+, including tools for editing, effects, face recognition, and geo-location.
- YouTube is a service that allows us to upload, view, and share videos. This service was acquired by Google in 2006 by 1650 million euros.

Developer Tools and Services

- Google Developers: is a page with technical documentation to use and exploit the resources of Google advanced form.
- Google Code: is a repository or warehouse where host (serve) code to use or share it freely with others.
- Google Analytics: is a powerful and useful statistics and analytics service for websites. Using a tracking code that we insert into our site creates detailed reports with all types of visitor data.
- Google Webmaster Tools: is a set of tools for those who have a blog or website on the Internet. Check the status of indexing by the search engine to optimize our visibility. It offers several reports that help us understand the extent of our publications also receive notifications for any serious error.
- Google Fonts: is a service that offers several online sources for use on any website.
- Google URL Shortener: is an application that allows to shorten URL service and also offers statistics.
- Google Trends: this service provides tools to compare the popularity of different terms and locate trend map search.
- Google Insights: a tool for checking the popularity of one or more search terms on the Internet, using the Google database.
- Apart from all these products and services related to the web and its many features, Google continues to diversify its market and is currently working on products whose purpose remains make life easier for people.
- Google Glasses: a display device, a kind of augmented reality glasses that let us browse the Internet, take pictures, capture videos, etc.
- Driverless car is a Google project that aims to develop a technology that allows the marketing of driverless vehicles. They are working with legislators to testing autonomous vehicles on public roads and already have the approval of the project in two states in the United States: California and Nevada. Still unknown is when technology will be available in the market.
- Android Home: is a Google home automation technology, which will connect the full home

to the Internet. For example, this service can tell us the time of completion of a product in the refrigerator.

Cross-References

► Google Analytics

Further Reading

- Jarvis, J. (2009). *What would Google do?* New York: Harper Collins.
- Page, L., Brin, S., Motwani, R., Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. <http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>.
- Poundstone, W. (2012). *Are you smart enough to work at Google?* New York: Little, Brown and Company.
- Stross, R. (2009). *Planet Google: One Company's audacious plan to organize everything we know*. New York: Free Press.
- Vise, D., & Malseed, M. *The Google story*. London: Pan Macmillan.

improve the site's content organization, increase the interaction of a website, and learn the reasons why visitors leave a site without making any purchase.

This service is free of charge, a feature emphasized in its presentation of the November 11, 2005. For experts, this circumstance is to redefine the existing business model at the moment, so far based on a charge depending on the volume of traffic analyzed. The benefit for the Telecommunications Company is focused on its advertising products: AdWords and AdSense, with an annual turnover estimated at 20 billion dollars. However, it is important to note that although Google offers this free service – except for those companies whose sites have more than five million visits per month – is necessary for all clients to invest in the implementation of this system.

Features

The service offers the following features:

1. Monitoring of several websites. Through the account on the platform, the user of this service has multiple views so can refer to specific reports of a domain or subdomain.
2. Monitoring of a blog, MySpace, or Facebook pages. It is possible to use Google Analytics on sites that cannot change the code of the page (for example, on MySpace). In that case, it is recommend the use of third party widgets in order to set up this service for predefined templates of this kind of websites (Facebook, MySpace, WordPress).
3. Follow-up to visits of RSS feeds, so it is necessary to previously run such a service tracking code. Since most of the programs for reading RSS feeds or Atom cannot execute JavaScript code, Google Analytics does not count page views that are recorded through an RSS reader. In order to track those page views by this service, visitors must run a JavaScript file on Google's servers.
4. Compatible with other web analytics solutions. Google Analytics can be used with

Google Analytics

Natalia Abuín Vences¹ and Raquel Vinader Segura^{1,2}

¹Complutense University of Madrid, Madrid, Spain

²Rey Juan Carlos University, Fuenlabrada, Madrid, Spain

Google Analytics is a web statistics service offered by the North American Telecommunications giant Google since the beginning of 2006. It provides great useful information and allows all type of companies to get fast and easily data about their website traffic. Google Analytics not only let to measure sales and conversions, but it also offers statistics on how visitors use a website, how they contact it, and what to do so they keep visiting it. According to the company, it is a service aimed at executives, marketers, and web and content developers. It is useful to optimize online marketing campaigns in order to increase their effectiveness,

other internal or third parties solutions that have been installed for this purpose. For a list of possible compatible solutions, please check its App Gallery.

The operation of Google Analytics is based on a Page Tag Technique, which consists in the collection of data through the browser of the visitor, which will be stored in data collector servers. This is a popular data collection method because it is technically easier and cheaper. Combine cookies, Internet browsers, and codes within each of the pages of the website. This information is collected through a JavaScript code (known as tags or beacons) generated by logging into Google Analytics. This fragment of JavaScript code (several in case of making different trackings for several domains) must be copied and inserted into the header of the source code of the website. When a user visits a specific page that contains this tracking code, it will be charged simultaneously to the other elements of the site and will generate a cookie: a small text message that a web server transmit to web browser to track user's activity on a website until the end of the visit. While this happens, all data captured will be loading Google servers that will turn on the Google Analytics panel by means of the corresponding reports will be available immediately.

Thus, this system allows obtaining real-time reports and, at the same time, having the possibility to be customized based on each client's interests. In addition, it offers the options of advanced segmentation and traffic flow view, that is, analyze the path of a visitor on a site, while it allows you to visually evaluate how users interact with the pages.

Metrics

The reports offered by Google Analytics on a certain web page provide the following information:

1. Visits

- 1.1. Visits received by the site. A visit is an interaction, by an individual, with a website.

- 1.2. Page views. Number of times a web page is loaded completely from the web server to the browser of a user during a given period.
- 1.3. Pages/visit. Pages per visit ratio. Measure the depth of the visit and is intimately related to the time spent on the website.
- 1.4. Bounce Rate or percentage of rebound. It is the percentage of visits only consulting a page of a site before you leave it.
- 1.5. Average time on site (average dwell time). It is the time means that visitors to a site are interacting with it. Serves to understand if users are engaged with the site. It is calculated by the difference between last and first page view to visitor request, not when users leave the page.
- 1.6. % new visits. Percentage of new visitors in the page.
2. Traffic sources
 - 2.1. Direct Traffic. Visitors arrive directly to the website by typing in your browser the address of the site.
 - 2.2. Referring sites or pages from which visitors arrive at the site measured.
 - 2.3. Search engines. Visits to the page from search engines are.
3. Place. Origin location of the visitor
 - 3.1. Number of visits to places of origin
 - 3.2. Graphic by region
 - 3.3. Map indicating places of visit.
4. Content: main contents of the site visited.

Other variables of measurement that can be obtained are the visitor language setting, the browser used, or the type of Internet connection used by the user.

Google Analytics displays these metrics both graphically and numerically helping the user to interpret main results.

Benefits of Google Analytics

Among others, we must mention the following:

1. Compatibility with third-party services. It offers the possibility of using applications

that improve data collected by Google Analytics. Many of these applications are contained in the App Gallery.

2. **Loading speed:** Google Analytics code is light and it is loaded asynchronously, so will not affect the loading of the page.
3. **Customized Dashboards.** Possibility to select abridged reports from the main sections of Google Analytics. Up to 12 different reports can be selected, modified, or added to the main page. These reports provide information in real time and can be activated to be distributed via email.
4. **Map Overlay Reports.** A graphical way of presenting data that reflect where around the world visitors are connecting from when viewing a website. This is possible thanks to IP address location databases and provide a clear representation of the parts of the world visitors are coming from.
5. **Data Export and Scheduling.** Report data can be manually exported in a variety of formats, including CSV, TSV, PDF, or the open-source XML.
6. **Multiple Language Interfaces and Support.** Google Analytics currently can display reports in 25 languages, and this number is growing continually.
7. In addition to Facebook and other social spaces monitoring, Google Analytics can track mobile websites and mobile applications on all web-enabled devices, whether or not the device runs JavaScript.

Some disadvantages of Google Analytics we can point out are:

- **Technical limitations:** in case of the visitor's browser does not have JavaScript code enabled or if some cookies capture functions are blocked, the results reported will not be accurate. This situation is not common but can occur and be relevant in sites with millions of hits.
- Some users claim that the interface is not as intuitive as it is often suggested. It usually requires a preparation time to get familiarized with the information offered. Google

developers work on the program to make it more intuitive.

- **Loss of data by mistake.** If the code provided by analytics and integrated in the source code of the site is deleted by mistake, i.e., while updating some themes in WordPress, any records registered during the time it is not working will be lost.

Privacy

As discussed previously, the service uses cookies to collect information about the interactions of certain website users. The reports offered by Google Analytics provide non-personally identifiable information as part of the Google policy and do not include any information about real IP address. That is, the information provided to Google's clients does not include any information which can be used to identify a user, such as personal names, addresses, emails, or credit cards numbers, among others. In fact, Google has implemented a privacy policy which is committed to keep the information stored in their computer systems in a secure manner.

Google Analytics protects the confidentiality of your data in different ways:

1. The clients of this service are not allowed to send personal information to Google.
2. Data are not shared without the user's consent, except in certain limited circumstances, such as disclosures required by law.
3. Investments in security systems. Engineering teams dedicated to security in Google fight against external data threats.

Certified Partner Network

Google Analytics has a network of certified partners around the world that provides support for the implementation of the system and offer expert analysis of the data collected. They are agencies and consultancies that provide applications of web statistics, analysis services and testing of

websites, and optimization services that have passed a long selection process.

To become a member of this network, it is necessary that interested companies overcome a training program that passes necessarily by obtaining Google Analytics Individual Qualification (GAIQ) certified by Google Analytics Conversion University.

A list of certified partners by country can be consulted on the Google Analytics site. Most important services provided are the following:

- Web measurement planning
- Basic, advanced, and custom implementations
- Mobile sites and applications development
- Technical assistance
- Google Analytics' API integrations
- Analysis and consulting for online media
- Online media channel allocation
- Websites and landing page testing
- Development of custom panels
- Training

At the same time, there is a program of authorized partners of Google Analytics Premium where the certified partners are ready to offer Google Analytics Premium package directly to customers.

Google Analytics Premium

The Premium service offers the same features as Google Analytics free version but includes extra features that make it an ideal tool for large companies. In exchange for a flat fee, offers a higher processing power for more detailed results, a dedicated team of service and support, warranties of service and up to thousands of millions of visits a month.

Google Analytics Premium allows the client to collect, analyze, and share more data than ever before.

- Extended data limits. Measure much more than before: up to one billion visits per month.
- Can use 50 custom variables, 10 times more than the standard version.

- Allows downloading without sample reports: up to three million rows of data without sample to analyze them with great precision.

This service also offers a dedicated account manager, who works as one member of the company team, and offers technical assistance in real time in order to ensure the quick resolution of incidents.

Google Analytics is a service that offers an enormous amount of data about the traffic on a particular website. Real importance lies in the necessary interpretation of this information to optimize processes, as well as designing and launching marketing campaigns.

Cross-References

- [Google](#)

Further Reading

- Clifton, B. (2010). *Advanced web metrics with Google analytics*. Indianapolis: Wiley.
- Google Inc. Google analytics support. <http://www.google.com/analytics/>. Accessed Aug 2014.
- Ledford, J., Teixeira, J., & Tyler, M. E. (2010). *Google analytics*. Indianapolis: Wiley.
- Quinonez, J. D. What is and how Google analytics. <http://www.whatsnew.com/2013/08/27/que-es-y-como-funciona-google-analytics>. Accessed Aug 2014.

Google Books Ngrams

Patrick Juola
Department of Mathematics and Computer Science, McNulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Synonyms

[Google Books Ngrams](#)

Introduction

The Google Books Ngram corpus is the largest publicly available collection of linguistic data in existence. Based on books scanned and collected as part of the Google Books Project, the Google Books Ngram Corpus lists the “word n -grams” (groups of 1–5 adjacent words, without regard to grammatical structure or completeness) along with the dates of their appearance and their frequencies, but not the specific books involved. This database provides information about relatively low-level lexical features in a database that is orders of magnitude larger than any other corpus available. It has been used for a variety of research tasks, including the quantitative study of lexicography, language change and evolution, human culture, history, and many other branches of the humanities.

Google Books Ngrams

The controversial Google Books project was an ambitious undertaking to digitize the world’s collection of print books. Google used high-speed scanners in conjunction with optical character recognition (OCR) to produce a machine-readable representation of the text contained in millions of books, mostly through collaboration with libraries worldwide. Although plagued by litigation, Google managed to scan roughly 25 million works between 2002 and 2015, roughly a fifth of the 130 million books published in human history (Somers, 2017).

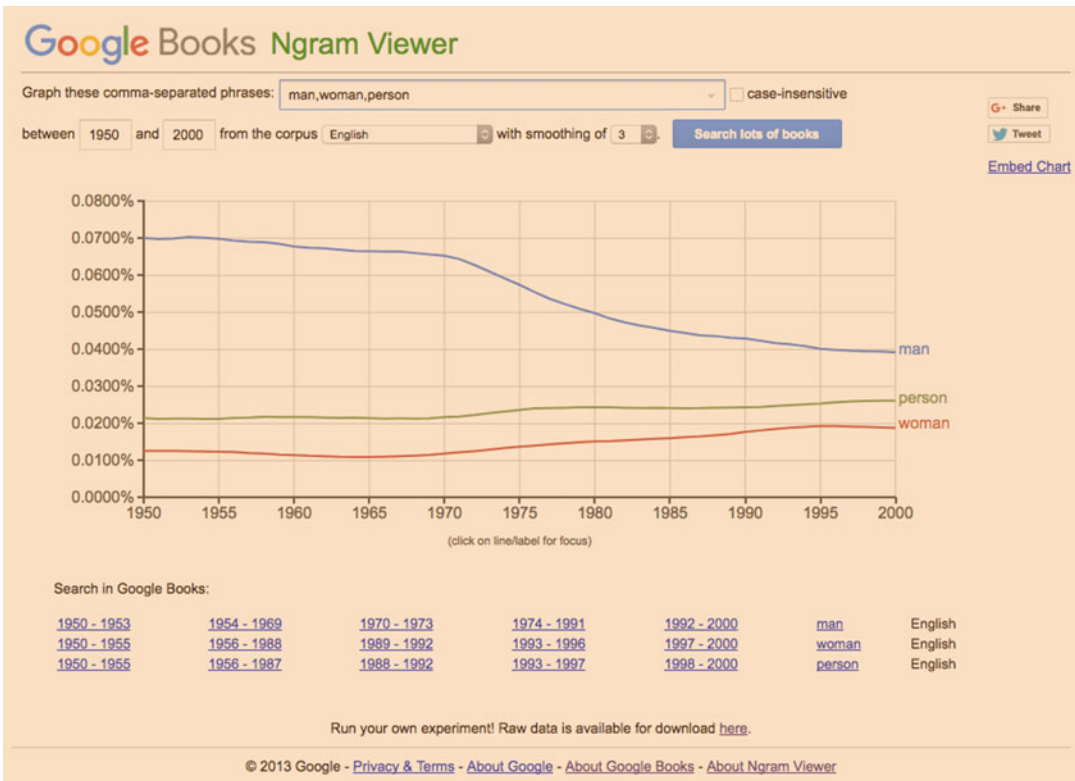
In part to satisfy the demands of copyright law, Google does not typically make the full text of books available, but is allowed to publish and distribute snippets. The Google Books Ngram corpus (Michel et al., 2011) provides n -grams (groups of n consecutive nonblank characters, separated by whitespace) for five million books at values of n from 1 to 5. A typical 1-gram is just a word, but could also be a typing mistake (e.g., “hte”), an acronym (“NORAD”), or a number (“3.1416”). A typical 2-gram would be a two-

word phrase, like “in the” or “grocery store,” while a typical 5-gram would be a five-word phrase like “the State of New York” or “Yesterday I drove to the.” These data are tabulated by year of publication and frequency to create the database. For example, the 1-gram “apple” appeared seven times in one book in 1515, 26 times across 16 books in 1750, and 24,689 times across 7871 books in 1966.

Google Books Ngrams provides data for several different major languages, including English, Chinese (simplified), French, German, Hebrew, Italian, Russian, and Spanish. Although there are relatively few books from the early years (only one English book was digitized from 1515), there are more than 200,000 books from 2008 alone in the collection, representing nearly 20 billion words published in that year. The English corpus as a whole has more than 350 billion words (Michel et al. 2011), substantially larger than other large corpora such as News on the Web (4.76 billion words), Wikipedia (1.9 billion words), Hansard (1.6 billion words), or the British National Corpus (100 million words), but does not provide large samples or contextual information.

Using the Google Books Ngrams Corpus

Google provides web access through a form, the Ngram Viewer, at <https://books.google.com/ngrams>. Users can type the phrases that interest them into the form, choose the specific corpus, and select the time period of interest. A sample screen shot is attached as Fig. 1. This plots the frequency of the words (technically, 1-grams) “man,” “woman,” and “person” from 1950 to 2000. Three relatively clear trends are visible in this figure. The frequency of “man” is decreasing across the time period, a drop-off that accelerates after 1970, while the frequency of both “woman” and “person” start to increase after 1970. This could be read as evidence of the growing influence of the feminist movement and its push towards more gender-inclusive language.



Google Books Ngrams, Fig. 1 Google Books Ngram Viewer (screen shot)

The Ngram Viewer provides several advanced features, like the ability to search for wildcards (“*”) such as “University of *.” Searching for “University of *” gives a list of the ten most common/popular words to complete that phrase. Perhaps as expected, this list is dominated in the 1700s and 1800s by the universities of Oxford and Cambridge, but by the early 1900s, the most frequent uses are the universities of California and Chicago, reflecting the increasing influence of US universities. The viewer can also search for specific parts of speech (for example, searching for “book_VERB” and/or “book_NOUN”), groups of inflectionally related words (such as “book,” “booking,” “books,” and “booked”), or even related in terms of dependency structure. In addition, the raw data is available for download from <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html> for people to use in running their own experiments.

Uses of the Corpus

The Google Books Ngram corpus has proven to be a widely useful research tool in a variety of fields. Michel et al. (2011) were able to show several findings, including the fact that the English lexicon is much larger than estimated by any dictionary; that there is a strong and measurable tendency for verbs to regularize (that is, for regular forms like “learned” or “burned” to replace irregular forms like “learnt” or “burnt”); that there is an increasingly rapid uptake of new technology into our discussions, and similarly, an increasing abandonment of old concepts such as the names of formerly famous people; and, finally, showed an effective method of detecting censorship and suppression. Other researchers have used this corpus to detect semantic shifts over time, to examine the cultural consequences of the shift from an urban to rural environment, to study the

growth of individualization as a concept, and to measure the amount of information in human culture (Juola, 2012). The Ngram corpus has been useful to establish how common a collocation is, and by extension, help assess the creativity of a proposed trademark. Both for small-scale (at the level of words and collocations) and large-scale (at the level of language or culture itself) investigations, the Ngram corpus creates new affordances for many types of research.

Criticisms of the Google Books Ngram Corpus

The corpus has been sharply criticized for several perceived failings, largely related to the methods of collection and processing (Pechenick et al., 2015). The most notable and common criticism is the overreliance on OCR. Any OCR process will naturally contain errors (typically 1–2% of the characters will be mis-scanned with high-quality images), but the error rates are much higher with older documents. One aspect that has been singled out for criticism (Zhang, 2015) is the “long s” or “medial s,” a pre-1800 style of printing the letter “s” in the middle of a word that looks very similar to the letter “f.” Figure 2 shows an example of this from the US Bill of Rights. Especially in the earlier (2009) version of the Google Books Ngram corpus, the OCR engine was not particularly good at distinguishing between the two, so the word “son” could be read as “fon” in images of early books.

Another issue is the representativeness of the corpus. In the early years of printing, book

publishing was a rare event, and not all books from that period survive. The corpus contains a single book from 1505, none at all from 1506, and one each from 1507, 1515, 1520, 1524, and 1525. Not until 1579 (three books) is there more than one book from a single year, and still, in 1654, only two books are included. Even as late as 1800, there are only 669 books in the corpus. Of course, the books that have survived to be digitized are the books that were considered worth curating in the intervening centuries, and probably do not accurately represent language as typically used or spoken.

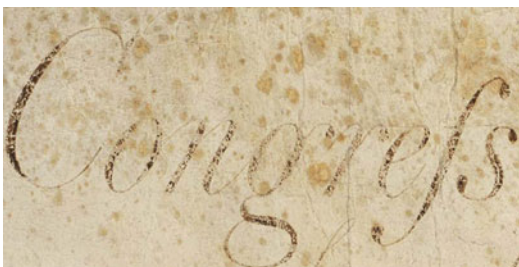
More subtly, researchers like Pechenick et al. (2015) have shown that the corpus itself is not well balanced. Each book is equally weighted, meaning that a hugely influential book like *Gulliver's Travels* (1726) has less weight on the statistics than a series of largely unread sermons by a prolific eighteenth-century minister that happened to survive. Furthermore, the composition of the corpus changes radically over time. Over the twentieth century, scientific literature starts to become massively overrepresented in the corpus, while fiction decreases, despite being possibly a more accurate guide to speech and culture. (The English corpus offers fiction-only as an option, but the other languages do not.)

Conclusion

Despite these criticisms, the Google Books Ngram corpus has proven to be an easily accessible, powerful, and widely useful tool for linguistic and cultural analysis. At more than 50 times the next largest linguistic data set, the Ngram corpus, provides access to more raw data than any other corpus currently extant. While not without its flaws, these flaws are largely inherent to any large corpus that relies on large-scale collection of text.

Cross-References

► [Corpus Linguistics](#)



Google Books Ngrams, Fig. 2 Example of “medial s” from the United States Bill of Rights (1788)

Further Reading

- Juola, P. (2013). Using the Google N-Gram corpus to measure cultural complexity. *Literary and linguistic computing*, 28(4), 668–675.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.
- Pechenick, E. A., Danforth, C. M., & Dodds, P. A. (2015). Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS One*, 10(10), e0137041. <https://doi.org/10.1371/journal.pone.0137041>.
- Somers, J. (2017). Torching the modern-day library of Alexandria. In *The Atlantic*. <https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/>. Accessed 23 July 2017.
- Zhang, S. (2015). The pitfalls of using google ngram to study language. In *Wired*. <https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/>. Accessed 23 July 2017.

Google Flu

Kim Lacey
Saginaw Valley State University, University
Center, MI, USA

Google Flu Trend (GFT) is a tool designed by Google to collect users' web searches to predict outbreaks of influenza. These trends are identified by tracking search terms related to symptoms of the virus combined with the geographic location of users. In terms of big data collection, GFT is seen as a success due to its innovative utilization of large amounts of crowd-sourced information. However, it has also been deemed somewhat a failure due to the misunderstanding of the symptoms of the flu and media-influenced searches. Matthew Mohebbi and Jeremy Ginsberg created GFT in 2008. As of 2014, GFT actively monitors 29 countries.

When GFT was first launched, reviewers praised its accuracy; in fact, GFT was touted to be 97% accurate. However, these numbers were discovered to be quite misleading. GFT analyzes Google searches to map geographic trends of illness. When an individual feels an illness coming on, one of the common, contemporary reactions is to

investigate these symptoms online. Such searches might be accurate (i.e., people searching for the correct affliction), but many are not. As such, this results in the high number of physicians noting patients are coming to a doctor's visit with a lists of symptoms and possible conditions, all discovered on websites by the likes of WebMD or Mayo Clinic. To further complicate the accuracy of GFT, many identify "cold" symptoms (e.g., runny nose) as flu symptoms, while in fact influenza is mainly a respiratory infection; the major symptoms include cough and chest congestion. One of the complications of GFT results from the algorithm for GFT not being accurately designed to "flag" the correct symptoms. In a separate study, David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani suggest that heavy media reporting of flu outbreaks led to many unnecessary Google searches about the flu. Because of this influx, GFT results spiked, falsely indicating the number of cases of the flu.

The combination of searching for incorrect flu symptoms and heavy media attention proved catastrophic for GFT's initial year of reporting. The inaccuracies actually went unnoticed until well after the 2008–2009 flu season. In 2009, an outbreak of H1N1 (popularly known as the swine flu) during the off-season also caused unexpected disruptions of GFT. Because GFT was designed to identify more common strains of influenza, the unpredicted H1N1 outbreak wreaked havoc on reporting trends. Even though H1N1 symptoms do not vary too much from common flu symptoms (it is the severity of the symptoms that causes the majority of the worry), its appearance during the off-season threw trackers of GFT for a loop. Additionally, H1N1 proved to be a global outbreak and thus another complication in the ability to accurately report the flu trends.

Within the first 2 years of GFT, it received a great amount of attention, for good and for ill. Some of the positive attention focused specifically on the triumph of big data. To put this phenomenon in perspective, Viktor Mayer-Schonberger and Kenneth Cukier point out that GFT utilizes billions of data points to identify outbreaks. These data points, which are collected from smaller sample groups, are then applied to larger populations to predict outbreaks. Mayer-Schonberger and

Cukier also note that before GFT, flu trends were always a week or so offtrack – in a sense, we were always playing catch-up. Google, on the other hand, recognized an opportunity. If Google could flag specific key word searches, both by volume and geographic location, it might be able to identify a flu outbreak as it was happening or even before it occurred. On the surface, this idea seems to be an obvious and positive use of big data. Diving deeper, GFT has received a lot of critical attention due to skewed reporting, incorrect algorithms, and misinformed interpretation of the data sets.

Mayer-Schonberger and Cukier imply that correlation has a lot to do with the success and failure of GFT. For example, the more people who search for symptoms of the flu in a specific location, the more likely there is to be an outbreak in that area. (In a similar vein, we might be able to use Google analytics to track how many people are experiencing a different event specific to a geographic location, such as a drought or a flood.) However, what Google did not take into consideration was the influence the media would have on its product. Once GFT began receiving media attention, more users began searching for both additional information on the Google Flu project and also flu-related symptoms. These searches were discovered to have arisen only because users had heard about GFT on the news or in other reporting. Because the GFT algorithm did not take into consideration traffic from media attention, the algorithm did not effectively account for the lack of correlation between searches related to media coverage and searches related to actual symptoms. To this day, Google's data sets remain at the unstable and struggle with the relation between actual cases of the flu and media coverage. As of yet, GFT does not have an algorithm that has been effectively designed to differentiate between media coverage and influenza outbreaks.

But still, there have been many researchers who have touted GFT as a triumph of collective intelligence, a signal that big data is performing in the ways researchers and academics imagined and hoped it would from the start. The ability to use large data sets is an impressive, and advantageous, use of mundane actions to establish health

patterns and prepare for outbreaks. Even though all users agree to Google's terms of service any time they utilize its search engines, few recognize what happens to this information beyond the returned search results. The returned results are not the only information that is being shared during a web search. And even though on its privacy policy page, Google acknowledges the ways it collects user data, what it collects, and what it will do to secure the privacy of that information, some critics feel this exchange is unfair. In fact, users share much more than they realize, thus the ability for Google to create a project such as GFT based on search terms and geolocation. The seemingly harmless exchange of search terms for search results is precisely from where Google draws its collection of data for GFT. By aggregating the loads of metadata users provide, Google hoped it would be able to predict global flu trends.

While all this data collection for predictive health purposes sounds great, there has been a lot of hesitation regarding the ownership and use of private information. For one, GFT has received some criticism because of the fact that it represents the shift in the access of data from academics to companies. Put simply, rather than academics collecting, analyzing, and sharing the information they glean from years of complex research, putting these data sets in the hands of a large corporation (such as Google) causes pause for users. In fact, for many users, the costs of health care and access to health providers leave few alternatives to web searches to find information about symptoms, preventative measures, and care. Another concern about GFT is the long-term effects of collecting large amounts of data. Because big data collection is fairly new, we do not know the ramifications of collecting information to which individuals do not have easy access. Even though Google's privacy policy states that user information will remain encrypted, the ways in which this information can be used and shared remain vague. This hesitancy is not exclusive to GFT. For example, in another form of big data, albeit a more personalized version, many believe the results of DNA sequencing being shared with insurance companies will lead to a similar loss of control over personal data. Even though individuals

should maintain legal ownership over personal health, the question of what happens when it merges with big data collection remains unclear. Further, the larger implications of big data are unknown, but projects like GFT and DNA sequencing pose the question of who owns our personal health data. If we do not have access to our own health information, or if we do not feel able to freely search for health-related issues, then Google's tracking might post more problems than it was designed to handle.

Along these lines, one of the larger concerns with GFT is the use of user data for public purposes. Once again echoing concerns about DNA sequencing, a critique of GFT is how Google collects its data and what it will do with it once a forecast has been made. Because Google's policies state that it may share user information with trusted affiliates or forcible government requests, further, some are worried that Google's collection of health-related data might lead to geographically specific ramifications (e.g., higher health insurance premiums). On the flip side, Miguel Helft, writing in *The New York Times*, notes that while some users are concerned about privacy issues, GFT did not alter any of Google's regular tracking devices, but instead allows users to become aware of flu trends in their area. Helft points out that Google is only using the data it originally set out to collect and has not adjusted or changed the ways it collects user information. This explanation, however, does not appease everyone, as some are still concerned with Google's lack of transparency in the process of collecting data. For example, GFT does not explain how the data is collected nor does it explain how the data will be used. Google's broadly constructed guidelines and privacy policy are, to some, flexible enough to apply to many currently unimagined intentions.

In response to many of these concerns, Google attempted to adjust the algorithm once again. Unfortunately, once again the change did not consider (or consider enough) the high media coverage GFT would receive. This time, by adjusting the algorithm on the assumption that media coverage was to blame for the GFT spike, it only

created more media coverage of the problems themselves, thus adding to (or at minimum sustaining) the spike in flu-related searches. At one point, Lazer, Kennedy, King, and Vespignani sought to apply the newly adjusted algorithm to previously collected data. To better evaluate the flu trends from the start of GFT, Lazer, Kennedy, King, and Vespignani applied the adjusted algorithm to backdated data found using the Wayback Machine, although trends were still difficult to identify because of the uncertainty of the influence of media coverage.

Another one of the more troubling issues with GFT is that it is failing to do what it was designed to do: forecast the Centers for Disease Control's (CDC) results. The disconnect between what actually occurred (how many people actually had the flu) and what GFT predicted was apparent in year 1. Critics of GFT are the fact that it consistently overshot the number of cases of the flu in almost every year since its inception in 2008. This overestimation is troubling not only because it signifies a misunderstanding of big data sets, but it also could potentially cause a misuse of capital and human resources for research. The higher the predicted number of flu cases, the greater amount of attention fighting that outbreak will receive. If these numbers remain high, financial resources will not be spent on projects which deserve more attention. David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani reviewed Google's algorithm and discovered in the 2011–2012 flu season alone (3 years into the project) GFT overestimated the number of cases by as much as 50% more than what the CDC reported. The following year, after Google retooled its algorithm, GFT still overestimated the number of flu cases by approximately 30%. Additionally, Lazer, Kennedy, King, and Vespignani noticed GFT estimates were high in 100 of the 108 weeks they tracked. The same four scientists suggested that Google was guilty of what they call "big data hubris": the assumption that big data sets are more accurate than traditional data collection and analysis. Further, the team suggested that GFT couples its data with CDC data. Since GFT was not designed as a substitute for doctor visits, by linking the predictions with

reported cases, GFT would be able to more effectively and accurately predict flu outbreaks.

GFT is not disappearing, however. It is a project that many are still behind because of its potential to impact global flu outbreaks. Google continues to back its trend analysis system because of its potential impact to recognize outbreaks of the flu and, eventually other, more serious diseases. One of the ways they are addressing concerns is by using what they call “nowcasting”: using data trends to provide daily updates rather than larger, seasonal predictions. Others, too, remain cautiously optimistic. Eric Topol suggests that while GFT is fraught with complications, the idea that big data collection could be applied to many different conditions is what we need to focus on.

Cross-References

- [Bioinformatics](#)
- [Biosurveillance](#)
- [Correlation Versus Causation](#)
- [Data Mining Algorithms](#)
- [Google](#)

Further Reading

- Bilton, N. Disruptions: Data without context tells a misleading story. *The New York Times: Bits Blog*. http://bits.blogs.nytimes.com/2013/02/24/disruptions-google-flu-trends-shows-problems-of-big-data-without-context/?_php=true&_type=blogs&_r=0. Accessed 27 Aug 2014.
- Blog.google.org. *Official google.org Blog: Flu Trends Updates Model to Help Estimate Flu Levels in the US*. <http://blog.google.org/2013/10/flu-trends-updates-model-to-help.html>. Accessed 27 Aug 2014.
- Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing Google Flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PloS One*, 6(8), e23610 <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0023610>. Accessed 27 Aug. 2014.
- Google.com. *Privacy Policy–Privacy & Terms–Google*. Available at: <http://www.google.com/intl/en-US/policies/privacy/#infosecurity>. Accessed 27 Aug 2014.
- Helft, M. Is there a privacy risk in Google Flu trends? *The New York Times: Bits Blog*. http://bits.blogs.nytimes.com/2008/11/13/does-google-flu-trends-raises-new-privacy-risks/?_php=true&_type=blogs&_r=0. Accessed 26 Aug 2014.

[com/2008/11/13/does-google-flu-trends-raises-new-privacy-risks/?_php=true&_type=blogs&_r=0](http://bits.blogs.nytimes.com/2008/11/13/does-google-flu-trends-raises-new-privacy-risks/?_php=true&_type=blogs&_r=0).

Accessed 26 Aug 2014.

- Lazer, D., Kennedy R., King G., & Vespignani A. The parable of Google Flu: Traps in big data analysis. *Science* [online] 343(6176), pp.1203–1205. Available at: <http://www.sciencemag.org/content/343/6176/1203.full>. Accessed 27 Aug 2014.
- Lazer, D., Kennedy R., King G., & Vespignani A. Google Flu still appears sick: An evaluation of the 2013–2014 Flu season. Available at: <http://gking.harvard.edu/publications/google-flu-trends-still-appears-sick%C2%A0evaluation-2013%E2%80%902014-flu-season>. Accessed 26 Aug 2014.
- Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin.
- Salzberg, Steven. Why Google Flu is a failure. *Forbes*. Available at: <http://www.forbes.com/sites/steven-salzberg/2014/03/23/why-google-flu-is-a-failure/>. Accessed 27 Aug 2014.
- Topol, E., Hill, D., & Tantor Media. (2012). *The creative destruction of medicine: How the digital revolution will create better health care*. New York: Basic Books.

Governance

Sergei A. Samoilenko¹ and Marina Shilina²

¹George Mason University, Fairfax, VA, USA

²Moscow State University (Russia), Moscow, Russia

The Impact of Big Data

The rise of Web 2.0 in the new millennium has drastically changed former approaches of information management. New social media applications, cloud computing, and software-as-a-service applications further contributed to the data explosion. The McKinsey Global Institute (2011) estimates that data volume is growing 40% per year and will continue to grow 44 times between 2009 and 2020. Primarily, the interactive data poses new challenges to enterprises that now have to deal with issues related to data quality and information life-cycle management. Companies constantly seek new ideas to better understand how to collect, store, analyze,

and use big data in ways that are meaningful to them.

Big data is generally referred within the context of “the three Vs” – *volume*, *velocity*, and *variety*. However, its polystructured nature made it necessary to review big data in the context of *value* or ways of utilizing all kinds of data including database content, log files, or web pages in a cost-effective manner. In many ways, both *internal data* (e.g., enterprise application data) and *external data* (e.g., web data) have become a core business asset for an enterprise. Most organizations now recognize big data as an enterprise asset with financial value. They often use big data for predictive analytics to improve business results.

Big data has the potential to add value across all industry segments. For example, collecting sensor data through in-home health-care monitoring devices can help analyze elderly patients’ health and vital statistics proactively. Health-care companies and medical insurance companies can then make time interventions to save lives or prevent expenses by reducing hospital admissions costs. In finances capital markets generate large quantities of stock market and banking transaction data that can help detect frauds and maximize successful trades. Electronic sensors attached to machinery, oil pipelines, and equipment generate streams of incoming data that can be used for preventive means to avoid disastrous failures. Streaming media, smartphones, and other GPS devices offer advertisers an opportunity to target consumers when they are in close proximity to a store or a restaurant.

Why Big Data Governance

Big data governance is a part of a broader information governance program that manages policies relating to data optimization, privacy, and monetization. A report from the Institute for Health Technology Transformation demonstrates that a standardized format for data governance is essential for health-care organizations to leverage the power of big data.

The process of governance refers to profiling the data, understanding what it will be used for, and then determining the required level of data management and protection. In other words, information governance is the set of principles, policies, and processes that corresponds to corporate strategy and define its operational and financial goals. These processes may include (a) following document policies relating to data quality, metadata, privacy, and information life-cycle management, (b) assigning new roles and responsibilities such as data stewards for improving the quality of customer data, (c) monitoring compliance with data policies regulating the work of customer service and call center agents, and (d) managing various data issues such as storing and eliminating duplicate records.

Big data governance determines what data is held, how it is held, where, and in what quality. According to Soares (2013), big data can be classified into five distinct types such as web and social media, machine to machine, big transaction data, biometrics, and human generated. Organizations need to establish the appropriate policies to prevent the misuse of big data and assess the reputational and legal risks involved when handling various data. For example, a big data governance policy might state that an organization will not integrate a customer’s Facebook profile into his or her master data record without that customer’s informed consent.

Big Data Governance Management

A prerequisite to efficient data governance is proper data management. In order to minimize potential risks related to data misuse or privacy violation, a strong information management should include a comprehensive data model supporting an enterprise’s business applications, proper data management tools, and methodology, as well as competent data specialists. A good data governance program assures adhering to the privacy, security, financial standards, and legal requirements. With effective information

governance in place, business stakeholders tend to have greater trust and confidence in data.

According to Mohanty et al. (2013), every enterprise needs an ecosystem of business applications, data platforms to store and manage the data, and reporting solutions. The authors discuss an Enterprise Information Management (EIM) framework that allows companies to meet the information needs of their stakeholders in compliance with appropriate organizational policies. The first component of EIM is selecting the right *business model*. There are three types of organization models: “decentralized model,” “shared services model,” and “independent model.” The “decentralized” model enables rapid analysis and execution outcomes produced by separate analytics teams in various departments. At the same time, generated insights are restrictive to a particular business function with little vision to strategic planning for an entire organization. The “shared services” model brings the analytics groups under a centralized management that potentially slows down insight generation and decision-making. The “independent” model has direct executive-level reporting and can quickly streamline requirements, however often lacks specific insights from each department.

Next, an EIM program needs to assure efficient *information management usage* that data and content are managed properly and efficiently and create a reference architecture that integrates emerging technologies into their infrastructure. Business requirements and priorities often dictate which *enterprise technology and architecture* to follow. For example, if the company decides they would like to interact with their customers through mobile channels, then the enterprise technologies and architectures will need to make provisions for mobility. EIM helps establish a data-driven *organization and culture* and introduce new roles such as data stewards within the enterprise. The company’s business priorities and road maps serve as a critical input to define what kind of *business applications* need to be built and when. For example, Apache Hadoop ecosystem can be used for distributing very large data files across all the nodes of a very large grid of servers in a way that supports recovery from the failure of

any node. Next, EIM helps in defining policies, standards, and procedures to find appropriate *data models and data stores* in an enterprise setup. For example, too many data models and data stores can cause severe challenges to the enterprise IT infrastructure and make it inefficient. *Information life-cycle management* is another process to monitor the use of information by data officers through its lifecycle, from creation through disposal, including compliance with legal, regulatory, and privacy requirements. Finally, EIM helps an organization estimate and address the regulatory risk that goes with *data regulations and compliance*. This helps some industries like financial services and health care in meeting regulatory requirements, which are of highest importance.

Soares (2012) introduces the IBM Information Governance Council Maturity Model as a necessary framework to address “the current state and the desired future state of big data governance maturity” (p. 28). This model is comprised of four groupings containing 11 categories:

Goals are the anticipated business outcomes of the information governance program that focuses on reducing risk and costs and increasing value and revenues.

Enables include the areas of organizational structures and awareness, stewardship, data risk management, and policy.

Core disciplines include data management, information life-cycle management, and information security and privacy.

Finally, *supporting disciplines* include data architecture, classification and metadata, and audit information logging and reporting.

Big Data Core Disciplines

Big data governance programs should be maintained according to new policies regarding the acceptable use of cookies, tracking devices, and privacy regulations. Soares (2012) addresses seven core disciplines of data governance.

The information governance organization assures the efficient integration of big data to an organizational framework by identifying the

stakeholders in big data governance and assigning new roles and responsibilities.

According to McKinsey Global Institute (2011), one of the biggest obstacles for big data is a shortfall of skills. With the accelerated adoption of deep analytical techniques, a 60% shortfall is predicted by 2018. The big data analytical capabilities include statistics, spatial, semantics, interactive discovery, and visualization. Adoption of unstructured data from external sources and increased demand for managing big data in real time requires additional management functions.

According to Venkatasubramanian (2013), data governance team comprises three layers. The *executive layer* comprised of senior management members who oversee the data governance function and ensure the necessary funding. A chief data officer (CDO) at this level is responsible for generating more revenue or decreasing costs through the effective use of data. The *strategic layer* is responsible for setting data characteristics, standards, and policies for the entire organization. Compliance officers integrate regulatory compliance and information retention requirements and help determine audit schedules. The legal team assesses information risk and determines if information capture and deletion are legally defensible. Data scientists use statistical, mathematical, and predictive modeling to build algorithms in order to ensure that the organization effectively uses all data for its analytics. The *tactical layer* implements the assigned policies. Data stewards are required to assist data analysts approve authorization for external data for business use. Data analysts conduct real-time analytics and use visualization platforms according to specific data responsibilities such as processing master/transactional data, machine-generated data, social data, etc. According to Breakenridge (2012), public relations and communications professionals should also be engaged in the development of social media policies, training, and governance. These may include research or audit efforts to identify potential areas of concern related to their brand's social media properties. They should work with senior management to build the social media core team to identify additional company policies that need to be incorporated into the social media policy (i.e., code of ethics, IT and

computing policies, employee handbook, brand guidelines, etc.). It will develop a communications plan to introduce necessary training in policy enforcement for directors or managers and then the social media policies to the overall employee population.

The *big data metadata* discipline refers to an organization process of metadata that describes the other data characteristics such as its name, location, or value. The big data governance program integrates big data terms within the business glossary to define the use of technical terms and language within that enterprise. For example, the term "unique visitor" is a unit used to count individual users of a website. This important term may be used in click-stream analytics by organizations differently: either to measure unique visitors per month or per week. Also, organizations need to address data lineage and impact analysis to describe the state and condition of data as it goes through diverse application processes.

The introduction of new sources of external personal data can lead to sudden security breach due to malware in the external data source and other issues. This could happen due to lack of enterprise-wide data standards, minimal metadata management processes, inadequate data quality and data governance measures, unclear data archival policies, etc. A big data governance program needs to address two key practices related to the *big data security and privacy* discipline. First, it would need to address tools related to *data masking*. These tools are critical to de-identify sensitive information, such as birth dates, bank account numbers, or social security numbers. These tools use data encryption to convert plain text within a database into a format that is unreadable to outsiders. *Database monitoring* tools are especially useful when managing sensitive data. For example, call centers need to protect the privacy of callers when voice recordings contain sensitive information related to insurance, financial services, and health care. The Payment Card Industry (PCI) Security Standards Council suggests that organizations use technology to prevent the recording of sensitive data and securely delete sensitive data in call recordings after authorization.

The *data quality* discipline ensures that the data is valid, is accurate, and can be trusted. Traditionally, data quality concerns relate to deciding on the data quality benchmarks to ensure the data will be fit for its intended use. It also determines the measurement criteria for data quality such as validity, accuracy, timeliness, completeness, etc. It includes clear communication of responsibilities for the creation, use, security, documentation, and disposal of information.

The *business process integration* program identifies key business processes that require big data, as well as key policies to support the governance of big data. For example, in oil and gas industry, the big data governance program needs to establish policies around the retention period for sensor data such as temperature, flow, pressure, and salinity on an oil rig for the period of drilling and production.

Another discipline called *master data integration* refers to the process when organizations enrich their master data with additional insight from big data. For example, they might want to link social media sentiment analysis with master data to understand if a certain customer demographic is more favorably disposed to the company's products. The big data governance program needs to establish policies regarding the integration of big data into the master data management environment. It also seeks to organize customer data scattered across business systems throughout the enterprise. Each data has specific attributes, such as customer's contact information that need to be complete and valid. For example, if an organization decides to merge Facebook data with other data, it needs to be aware that they cannot use data on a person's friends outside of the context of the Facebook application. In addition, it needs to obtain explicit consent from the user before using any information other than basic account information such as name, e-mail, gender, birthday, current city, etc.

The components of a big data *life-cycle management* include (a) information archiving of structured and unstructured information, (b) maintenance of laws and regulations that determine a retention of how long documents should be

kept and when they should be destroyed, and (c) legal holds and evidence collection requiring companies to preserve potential evidence such as e-mail, instant messages, Microsoft Office documents, social media, etc.

Government Big Data Policies and Regulations

Communications service providers (CSPs) now have access to more complete data on network events, location, web traffic, channel clicks, and social media. Recently increased volume and types of biometric data require strict governance relating to privacy and data retention. Many CSPs actively seek how to monetize their location data by selling it to third parties or develop new services. However, big data should be used considering the ethical and legal concerns and associated risks.

For many years, the European Union has established a formalized system of privacy legislation, which is regarded as more rigorous than the one in the USA. Companies operating in the European Union are not allowed to send personal data to countries outside the European Economic Area unless there is a guarantee that it will receive adequate levels of protection at a country level or at an organizational level. According to the European Union legal framework, employers may only adopt geolocation technology when it is demonstrably necessary for a legitimate purpose, and the same goals cannot be achieved with less intrusive means. The European Union Article 29 Data Protection Working Party states that providers of geolocation applications or services should implement retention policies that ensure that geolocation data, or profiles derived from such data are deleted after a "justified" period of time. In other words, an employee must be able to turn off monitoring devices outside of work hours and must be shown how to do so.

In January of 2012, the European Commission came up with a single law, the General Data Protection Regulation (GDPR), which intended to unify data protection within the European Union (EU).

This major reform proposal is believed to become a law in 2015. A proposed set of consistent regulations across the European Union would protect Internet users from clandestine tracking and unauthorized personal data usage. This new legislation would consider the important aspects of globalization and the impact of social networks and cloud computing. The Data Protection Regulation will also hold companies accountable for various types of violations based on their harmful effect. In the USA, data protection law is comprised of a patchwork of federal and state laws and regulations, which govern the treatment of data across various industries and business operations. The US legislation has been more lenient with respect to web privacy. Normally, the Cable Act (47 USC § 551) and the Electronic Communications Privacy Act (18 USC § 2702) prohibit operators and telephone companies to offer telephony services without the consent of clients and also prevent disclosure of customer data, including location. When a person uses a smartphone to place a phone call to a business, that person's wireless company cannot disclose his or her location information to third parties without first getting express consent. However, when that same person uses that same phone to look that business on the Internet, the wireless company is legally free to disclose his or her location. While no generally applicable law exists, some federal laws govern privacy policies in specific circumstances, such as:

US-EU Safe Harbor is a streamlined process for US companies to comply with the EU Directive 95/46/EC on the protection of personal data. Intended for organizations within the EU or USA, the Safe Harbor Principles are designed to prevent accidental information disclosure or loss of customer data.

Children's Online Privacy Protection Act (COPPA) of 1998 affects websites that knowingly collect information about or target at children under the age of 13. Any such websites must post a privacy policy and adhere to enumerated information-sharing restrictions. Operators are required to take reasonable steps to ensure that children's personal information is disclosed only to service providers

and third parties capable of maintaining the confidentiality, security, and integrity of such information. The law requires businesses to have apps and websites directed at children to give parental notice and obtain consent before permitting third parties to collect children's personal information through plug-ins. At the same time, it only requires that personal information collected from children be retained only "as long as is reasonably necessary to fulfill the purpose for which the information was collected."

The Health Insurance Portability and Accountability Act (HIPAA) requires notice in writing of the privacy practices of health-care services. If someone posts a complaint on Twitter, the health plan might want to post a limited response and then move the conversation offline. The American Medical Association requires physicians to maintain appropriate boundaries within the patient-physician relationship according to professional ethical guidelines and separating personal and professional content online. Physicians should be cognizant of patient privacy and confidentiality and must refrain from online postings of identifiable patient information.

The Geolocation Privacy and Surveillance Act (GPS Act) introduced in the US Congress in 2011 seeks to state clear guidelines for government agencies, commercial entities, and private citizens pertaining to when and how geolocation information can be accessed and used. The bill requires government agencies to get a cause warrant to obtain geolocation information such as signals from mobile phones and global positioning system (GPS) devices. The GPS Act also prohibits businesses from disclosing geographical tracking data about its customers to others without the customers' permission.

The Genetic Information Act of 2008 prohibits discrimination in health coverage and employment based on genetic information. Although this act does not extend to life insurance, disability insurance, or long-term care insurance, most states also have specific laws that prohibit the use of genetic information in these contexts.

Collection departments may use customer information from social media sites to conduct “skip tracking” to get up-to-date contact information on a delinquent borrower. However, they have to adhere to regulations such as the US Fair Debt Collection Practices Act (FDCPA) to prevent collectors from harassing debtors or infringing on their privacy. Also, collectors would be prohibited from creating a false profile to friend a debtor on Facebook or tweeting about an individual’s debt.

Today, facial recognition technology enables the identification of an individual based on his or her facial characteristics publicly available on social networking sites. Facial recognition software with data mining algorithms and statistical re-identification techniques may be able to identify an individual’s name, location, interests, and even the first five digits of the individual’s social security number. The Federal Trade Commission (2012) offers recommendations for companies to disclose to consumers that the facial data they use might be used to link them to information from third parties or publicly available sources.

Some states have implemented more stringent regulations for privacy policies. The California Online Privacy Protection Act of 2003 – Business and Professions Code sections 22575–22579 – requires “any commercial web sites or online services that collect personal information on California residents through a web site to conspicuously post a privacy policy on the site.” According to Segupta (2013), in 2014 California passed three online privacy bills. One gives children the right to erase social media posts, another makes it a misdemeanor to publish identifiable nude pictures online without the subject’s permission, and a third requires companies to tell consumers whether they abide by “do not track” signals on web browsers. In 2014 Texas passed a bill introduced that requires warrants for e-mail searches, while Oklahoma enacted a law meant to protect the privacy of student data. At least three states proposed measures to regulate who inherits digital data, including Facebook passwords, when a user dies. In March 2012 Facebook released a

statement condemning employers for asking job candidates for their Facebook passwords. In April 2012, the state of Maryland passed a bill prohibiting employees from having to provide access to their social media content.

In 2014 the 11th US Circuit Court of Appeals issued a major opinion extending Fourth Amendment protection to cell phones even when searched incident to an arrest. Police need a warrant to track the cell phones of criminal suspects. Investigators must obtain a search warrant from a judge in order to obtain cell phone tower tracking data that is widely used as evidence to show suspects were in the vicinity of a crime. As such, obtaining the records without a search warrant is a violation of the Fourth Amendment’s ban on unreasonable searches and seizures.

According to Byers (2014), “while most mobile companies do have privacy policies, but they aren’t often communicated to users in a concise or standardized manner.” The National Telecommunications and Information Administration suggested a transparency blueprint designed in 2012 and 2013 that called for applications to clearly describe what kind of information (e.g., location, browser history, or biometric data) it collects and shares. While tech companies (e.g., Google and Facebook) have tried to be more clear about the data they collect and use, most companies still refuse to adhere to such a code of conduct due to increased liability concerns. A slow adoption for such guidelines is also partially due to the government’s failure to push technology companies to act upon ideas for policy change. In May 2014, the President’s Council of Advisors on Science and Technology (PCAST) released a new report, *Big Data: A Technological Perspective*, which details the technical aspects of big data and new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected. In addition to a number of recommendations related developing privacy-related technologies, the report recommends that Congress pass national data breach legislation, extend privacy protections to non-US citizens, and update the Electronic Communications Privacy Act, which controls how the government can access e-mail.

Further Reading

- Breakenridge, D. (2012). *Social media and public relations: Eight new practices for the PR professional*. New Jersey: FT Press.
- Byers, A. (2014). W.H.'s privacy effort for apps is stuck in neutral. *Politico*. p. 33.
- Federal Trade Commission. (1998). Children's online privacy protection rule ("COPPA"). Retrieved from <http://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule>.
- Federal Trade Commission. (2012). Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. Retrieved from <http://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers>.
- Institute for Health Technology Transformation. (2013). Transforming health care through big data strategies for leveraging big data in the health care industry. Retrieved from <http://ihealthtran.com/wordpress/2013/03/ihf%C2%B2-releases-big-data-research-report-download-today/>.
- McKinsey Global Institute. (2011, May). Big data: The next frontier for innovation, competition, and productivity. Retrieved from <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>.
- Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big data imperatives: Enterprise 'big data' warehouse, 'BI' implementations and analytics (the Expert's voice)*. New York: Apress.
- Segupta, S. (2013). No action in Congress, so states move to enact privacy laws. *Star Advertiser*. Retrieved from http://www.staradvertiser.com/news/20131031_No_Action_In_Congress_So_States_Move_To_Enact_Privacy_Laws.html?id=230001271.
- Soares, S. (2012). Big Data Governance. Information Asset, LLC.
- Soares, S. (2013). A Platform for Big Data Governance and Process Data Governance. Boise, ID: MC Press Online, LLC.
- The President's Council of Advisors on Science and Technology. (2014). Big data and privacy: A technological perspective. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.
- Venkatasubramanian, U. (2013). Data governance for big data systems [White paper]. Retrieved from http://www.lintinfotech.com/resources/documents/datagovernanceforbigdatasystems_whitepaper.pdf.

Governance Instrument

► Regulation

Granular Computing

Davide Ciucci

Università degli Studi di Milano-Bicocca, Milan, Italy

Introduction

Granular Computing (GrC) is a recent discipline that deals with representing and processing information in the form of *information granules* or simply granules that arise in the process of data abstraction and knowledge extraction from data. The concept *information granularity* was introduced by Zadeh in 1979 (Zadeh 1979); however, the term *granular computing* was coined by Lin in 1997 (Lin 1997), and in the same year it was used again by Zadeh in (Zadeh 1997).

According to Zadeh, an information granule is a chunk of knowledge made of different objects "drawn together by indistinguishability, similarity, proximity or functionality" (Zadeh 2008). A granule is related to uncertainty management in the sense that it represents a lack of knowledge on a variable X . Indeed, instead of assigning it a precise value u , we use a granule, representing "some information which constrains possible values of u " (Zadeh 2008).

GrC is meant to group under the same formal framework a set of techniques and tools exploiting abstraction for approximate reasoning, decision theory, data mining, machine learning, and the like.

At present, it is not yet a formalized theory with a unique methodology, but it can be rather viewed as a unifying discipline of different fields of research. Indeed, it includes or intersects interval analysis, rough set theory, fuzzy set theory, interactive computing, and formal concept analysis, among others (Pedrycz et al. 2008).

Granule and Level Definition

The main concepts of GrC are of course *granule* and *levels of granularity*, which are closely

related: a level is the collection of granules of similar nature. Each level gives a different point of view (sometimes called *granular perspective* (Keet 2008)) to the subject under investigation.

Just to make a simple and typical example, structured writing can be described through the GrC paradigm. An article or a book can be viewed at different levels of granularity, from top to bottom: the article (book) itself, chapters, sections, paragraphs, and sentences.

We can move from one level to another, by going from top to bottom by decomposing a whole into parts through a *refinement* process. Or the other way round, going to an upper level merging parts into wholes, by a *generalization* process. For instance, in an animal taxonomy, the category of felines can be split into tigers, cats, lions, etc, or in the opposite direction, Afghan Hound, Chow Chow, and Siberian Husky can all be seen as dogs at a more general (abstract) level.

Thus, according to the level of granularity taken into account, i.e., to the point of view, a granule “may be an element of another granule and is considered to be a part forming the other granule. It may also consist of a family of granules and is considered to be a whole” (Yao 2008).

Granulation can be characterized according to different dimensions: based on a scale or not-scale-dependent; the relationship between two levels; the focus being on the granules or on the levels; the mathematical representation. This leads to a taxonomy of types of granularity; for a detailed discussion on this point, we refer to Keet (2008).

The idea of granule is quite akin to that of cluster, that is the elements of the same granule should be related, whereas the elements of two different granules should be sufficiently different to be separated. Thus, it is clear that all clustering algorithms can be used to granulate the universe. In particular, hierarchical algorithms produce not only the granulation at a fixed level, but the whole hierarchical structure. Other typical tools to build granules from data arise in the computational intelligence field: rough sets, fuzzy sets, interval

computation, shadowed sets, and formal concept analysis. Many of these tools are also typical of knowledge representation in presence of uncertainty, as described in the following.

Rough set theory is a set of mathematical tools to represent imprecise and missing information and data mining tools to perform feature selection, rule induction, and classification. The idea of granulation is at the core of the theory, indeed, the starting point is a relation (typically, equivalence or similarity) used to group indiscernible or similar objects. This relation is defined on object features; thus, two objects are related if they have equal/similar values for the features under investigation. For instance, two patients are indiscernible if they have the same symptoms. The obtained granulation is thus a partition (in case of equivalence relation) or a covering (in case of a weaker relation) of the universe, based on the available knowledge, that is, the features under investigation (that may also contain missing values).

Fuzzy sets are a generalization of Boolean sets, where each object is associated with a membership degree (typically a value in $[0,1]$) to a given subset of the universe, representing the idea that membership can be total or partial. A linguistic variable is then defined as a collection of fuzzy sets describing the same variable. For example, the linguistic variable *Temperature* can have values low, medium, or high, and these last are defined as fuzzy sets on the range of temperature degrees. It turns out that these values (i.e., low, medium, and high) can be seen as *graduated* granular values. That is, medium temperature is not described by a unique and precise value but by a graduated collection of values (Zadeh 2008). Shadowed sets can be seen as a simplified and more computationally treatable form of fuzzy sets, where uncertainty is localized due to a formal criterion.

Interval computation is based on the idea that measurements are always imprecise. Thus, instead of representing a measurement with a single precise value, an interval is used. Intervals can also be the result of the discretization of a

continuous variable. A calculus with intervals is then needed to compute with intervals. In this approach, any interval is a granule. Thus, different granulations can differ in precision and scale. An important aspect with respect to discretization of continuous variables is the definition of the suitable granularity to adopt that should be specific enough to address the problem to handle and make the desired characteristics emerge, yet avoiding intractability due to a too high detail.

Formal concept analysis is a formal framework based on lattice theory, aimed to create a concept hierarchy (named concept lattice) from a set of objects, where each concept contains the objects sharing the same features. Hence, a concept can be viewed as a granule and the concept lattice as the hierarchy of levels.

Granular Computing and Big Data

GrC is a way to organize data at a suitable level of abstraction, by ignoring irrelevant details, to be further processed. This is useful to handle noise in data and to reduce the computational effort. Indeed, a granule becomes a single point at the upper level, thus simplifying the data volume. In this way, an approximate solution is obtained, which can be refined in a following step if needed. This idea is explored in (Slezak et al. 2018) by providing “an engine that produces high value approximate answers to SQL statements by utilizing granulated summaries of input data”. It is to be noticed that in order to cope with big data, the engine gives *approximate* answers, so one should be aware that velocity comes to the price of precision.

Moreover, the representation of granules in a hierarchy permits to represent and analyze the available data from different perspectives according to a different granulation (multiview) or at a different level of abstraction (multilevel). From a more general and philosophical perspective, this possibility is also seen as a way to reconcile reductionism with system theory, since it preserves the split of a whole into parts and an

emerging behavior at an upper level in the hierarchy (Yao 2008).

Conclusion

Granular computing is an emerging discipline aimed to represent and analyze data in form of chunk of knowledge, the granules, connected in a hierarchical structure, and it exploits abstraction to reach its goals. As such, it complies with the human way of thinking and knowledge organization. In big data, granular computing can be used to reduce volume and to provide different points of view on the same data.

Cross-References

- [Data Mining](#)
- [Ontologies](#)

Further Reading

- Keet, C. M.. (2008). A formal theory of granularity. Ph.D. thesis, KRDB Research Centre, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy.
- Lin, T. Y.. (1997). Granular computing: From rough sets and neighborhood systems to information granulation and computing in words. In *Proceedings of European congress on intelligent techniques and soft computing* (pp. 1602–1606). Aachen: Germany.
- Pedrycz, W., Skowron, A., & Kreinovich, V. (Eds.). (2008). *Handbook of granular computing*. Chichester: Wiley.
- Slezak, D., Glick, R., Betlinski, P., & Synak, P. (2018). A new approximate query engine based on intelligent capture and fast transformations of granulated data summaries. *Journal of Intelligent Information System*, 50(2), 385–414.
- Yao, Y. Y.. (2008). Granular computing: Past, present and future. In *Proceedings of 2008 IEEE international conference on granular computing*. Hangzhou: China
- Zadeh, L. (1979). Fuzzy sets and information granularity. In N. Gupta, R. Ragade, & R. Yager (Eds.), *Advances in fuzzy set theory and applications* (pp. 3–18). Amsterdam: North-Holland.
- Zadeh, L. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 90(2), 111–127.
- Zadeh, L. (2008). Is there a need for fuzzy logic? *Information Sciences*, 178, 2751–2779.

Graph-Theoretic Computations/Graph Databases

John A. Miller¹, Arash Jalal Zadeh Fard^{1,2} and Lakshmish Ramaswamy¹

¹Department of Computer Science, University of Georgia, Athens, GA, USA

²Vertica (Hewlett Packard Enterprise), Cambridge, MA, USA

Introduction

The new millennium has seen a very dramatic increase in applications using massive datasets that can be organized in the form of graphs. Examples include Facebook, LinkedIn, Twitter, and ResearchGate. Consequently, a branch of big data analytics called *graph analytics* has become an important field for both theory and practice. Although foundations come from graph theory and graph algorithms, graph analytics focuses on computations on large graphs, often to find interesting paths or patterns.

Graphs come in two flavors, undirected and directed. One may think of an undirected graph as having locations connected with two-way streets and a directed graph having one-way streets. As directed graphs support more precision of specification and can simulate the connectivity of undirected graphs by replacing each undirected edge $\{u, v\}$ with two directed edges (u, v) and (v, u) , this reference will focus on directed graphs.

More formally, a *directed graph* or *digraph* may be defined as a two-tuple $G(V, E)$ where we have:

$$\begin{aligned} V &= \text{set of vertices} \\ E &\subseteq V \times V \quad (\text{set of edges}) \end{aligned} \quad (1)$$

A directed edge $e \in E$ is an ordered pair of vertices $e = (u, v)$ where $u \in V$ and $v \in V$. Given a vertex u , the set $\{v \mid (u, v) \in E\}$ is referred to as the *children* of u . *Parents* can be defined by following the edges in the opposite

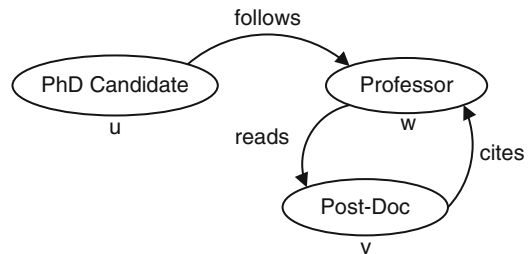
direction. Directed graphs are widely used in social networks. For example, V could be the set of registered users and E could represent a *follows* relationship say in ResearchGate.

In graph analytics and databases, vertices and/or edges often have labels attached to provide greater information content. Formally, labels may be added to a digraph through functions mapping vertices and/or edges to labels:

$$\begin{aligned} l_v : V &\rightarrow L_v && (\text{vertex labeling function}) \\ l_e : V &\rightarrow L_e && (\text{edge labeling function}) \end{aligned} \quad (2)$$

Many applications in graph analytics work with *multi-digraphs* that allow multiple edges from a vertex u to a vertex v , so long as the edge labels are distinct. As an example, let V be registered users in ResearchGate and E represents relationships between these users/researchers. Given three vertices u , v , and w , the vertices and edges could be labeled as follows: $l_v(u) = \text{"PhD Candidate"}$, $l_v(v) = \text{"Post - Doc"}$, $l_v(w) = \text{"Professor"}$, $l_e(u, w) = \text{"follows"}$, $l_e(v, w) = \text{"cites"}$, and $l_e(w, v) = \text{"reads"}$ (Fig. 1).

While graph theory can be traced back to work by Euler in the eighteenth century, work on graph algorithms began in earnest in the 1950s (e.g., Bellman-Ford-Moore algorithm and Dijkstra's algorithm for finding shortest paths). Over the years, a wide range of graph computations and algorithms have been developed. For big data graph analytics, work can be categorized into four main areas: Paths in



Graph-Theoretic Computations/Graph Databases, Fig. 1 An example graph in ResearchGate

Graphs, Graph Patterns, Graph Partitions, and Graph Databases.

Paths in Graphs

Many problems in graph analytics involve finding paths in large graphs; e.g., what is the connection between two users in a social networking application, or what is the shortest route from address A to address B.

An intuitive way to define path is to define it in terms of trail. A *trail* of length n in a digraph or multi-digraph can be defined as a sequence of non-repeating edges.

$$t = (e_1, \dots, e_n) \quad (3)$$

such that for all i , the consecutive edges e_i, e_{i+1} must be of the form (u, v) and (v, w) . A (simple) *path* p is then just a trail in which there are no repeating vertices. More specifically, $path(u, v)$ is a path beginning with vertex u and ending with vertex v .

The existence of a path from u to v means that v is *reachable* from u . In addition to finding a path from u to v , some applications may be interested in finding all paths (e.g., evidence gathering in an investigation) or a sufficient number of paths (e.g., reliability study or traffic flow).

As indicated, the length $len(path(u, v))$ is the number of edges in the path. The weighted length $wlen(path(u, v))$ is the sum of the edge labels as weights in the path.

A particular path $path_s(u, v)$ is a *shortest path* when $len(path_s(u, v))$ is the least among all paths from u to v (may also be defined in terms of $wlen$).

The position of a vertex within an undirected graph can also be defined in terms of paths. The *eccentricity* of a vertex $u \in V$ is defined as the length of the maximum shortest path from u to any other vertex:

$$ecc(u) = \max\{len(path_s(u, v)) | v \in V\} \quad (4)$$

Now, the *radius* of a (connected) graph is simply the minimum eccentricity, while the *diameter* of a graph is the maximum eccentricity. Although

eccentricity can be defined for digraphs, typically eccentricity, radius, and diameter are given in terms of its underlying undirected graph (directed edges turned into undirected edges).

Issues related to paths/connectivity include measures of influence in social media graphs. Measures of influence of a vertex, v (e.g., a Twitter user), include *indegree*, *outdegree*, and (undampened) *PageRank* (PR).

$$\begin{aligned} indegree(v) &= |\{u | (u, v) \in E\}| \\ outdegree(v) &= |\{w | (v, w) \in E\}| \\ PR(v) &= \sum \left\{ \frac{PR(u)}{outdegree(u)} | (u, v) \in E \right\} \end{aligned} \quad (5)$$

Graph Patterns

In simple terms, finding a pattern in a graph is finding a set of similar subgraphs in that graph. There are different models for defining similarity between two subgraphs, and we will introduce a few in this section. When unknown patterns need to be discovered (e.g., finding frequent subgraphs), it is called *graph pattern mining*. In comparison, when the pattern is known in advance and the goal is to find the set of its similar subgraphs, it is called *graph pattern matching*. In some applications of graph pattern matching, it is not the set of similar subgraphs that is important but its size. For example, counting the number of triangles in a graph is used in many applications of social networks (Tsourakakis et al. 2009).

The simplest form of pattern query is to take a query graph Q and match its labeled vertices to corresponding labeled vertices in a data graph G ; i.e., $pattern(Q, G)$ is represented by a multivalued function Φ :

$$\begin{aligned} \Phi : Q.V &\rightarrow 2^{G.V} s.t. \forall u' \in \Phi(u), l_v(u') \\ &= l_v(u) \end{aligned} \quad (6)$$

In addition to matching the labels of the vertices, patterns of connectivity should match as well. Common connectivity is established by examining edges (models may either ignore or take edge labels into account).

Traditional *graph similarity* matching can be grouped as *graph morphism* models. This group introduces complex and often quite constrained forms of pattern matching. The most famous models in this group are *graph homomorphism* and *subgraph isomorphism*.

- Graph homomorphism: It is a function f mapping each vertex $u \in Q.V$ to a vertex $f(u) \in G.V$, such that (1) $l_v(u) = l_v(f(u))$ and (2) if $(u, v) \in Q.E$, then $(f(u), f(v)) \in G.E$. For graph pattern matching, all or a sufficient number of graph homomorphisms can be retrieved.
- Subgraph isomorphism: It is a more restrictive form of graph homomorphism where we simply change the mapping function f to a bijection onto a subgraph of G .

High computational complexity and inability of these models to find certain meaningfully similar subgraphs in new applications have led to a more recently emerging group of graph pattern models called *simulation*. The major models in this group are *graph simulation*, *dual simulation*, *strong simulation*, *strict simulation*, *tight simulation*, and *CAR-tight simulation*.

- Graph simulation (Henzinger et al. 1995): Algorithms for finding graph simulation matches typically follow a simple approach. For each vertex $u \in Q.V$, initially compute the mapping set $\Phi(u)$ based on label matching. Then, repeatedly check the child match condition for all vertices to refine the mapping Φ until there is no change. The *child match condition* is simply that if $u' \in \Phi(u)$, then the labels of the children of u' that are themselves within Φ must include all the labels of the children of u .
- Dual simulation (Ma et al. 2011): It adds a *parent match condition* to graph simulation. The *parent match condition* is simply that if $u' \in \Phi(u)$, then the labels of the parents of u' that are themselves within Φ must include all the labels of the parents of u .

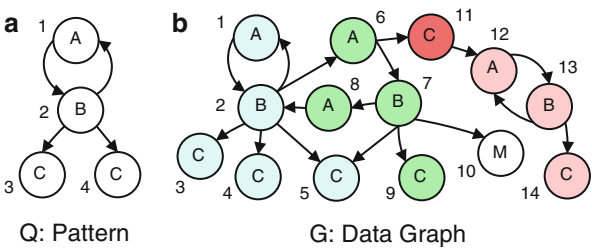
- Strong simulation (Ma et al. 2014): As dual simulation allows counterintuitive solutions that contain large cycles, various locality restrictions may be added to dual simulation to eliminate them. For strong simulation, any solution must fit inside a *ball* of radius equal to the diameter of the query graph Q .
- Strict simulation (Fard et al. 2013): Based on strong simulation, it applies dual simulation first to reduce the number of balls. Balls are only made from vertices that are in the image of Φ . This also reduces the number of solutions, making the results closer to those of traditional models like subgraph isomorphism.
- Tight simulation (Fard et al. 2014b): The solutions can be further tightened by reducing the number of balls and making them smaller. First a central vertex u_c ($ecc(u_c)$ equal to the radius) of the query graph Q is chosen, and then balls are created only for $u' \in \Phi(u_c)$. In addition, the radius of the balls is now equal to the radius of Q , not its diameter as before.
- CAR-tight simulation (Fard et al. 2014a): Results even closer to subgraph isomorphism can be obtained by adding a further restriction to tight simulation. A *cardinality restriction* on child and parent matches pushes results toward one-to-one correspondences. This modification is referred to as *cardinality restricted (CAR)-tight simulation*.

Figure 2 illustrates a simple example of subgraph pattern matching. For the given query Q (query graph), applying different pattern matching models on G (data graph) yields different results. In this figure, the numbers are the IDs of the vertices and the letters are their labels. Table 1 summarizes the results for different models where Φ is a multi-valued function (could be represented as a relation) and f is a function from vertices of Q to vertices of G ($f: Q.V \rightarrow G.V$). The table gives all such non-redundant functions/mappings. Subgraph pattern matching has applications in analyzing social

Graph-Theoretic Computations/Graph Databases,

Fig. 2 Example of subgraph pattern matching with different models

A: Arts Book
B: Biography Book
C: Children’s Book
M: Music CD



Graph-Theoretic Computations/Graph Databases, Table 1 Results of different pattern matching models of Fig. 2

Model	Subgraph results
Tight simulation	$\Phi(1, 2, 3, 4) \rightarrow (1, 2, \{3, 4, 5\}, \{3, 4, 5\}), (12, 13, 14, 14)$
CAR-tight simulation	$\Phi(1, 2, 3, 4) \rightarrow (1, 2, \{3, 4, 5\}, \{3, 4, 5\})$
Subgraph isomorphism	$f(1, 2, 3, 4) \rightarrow (1, 2, 3, 4), (1, 2, 3, 5), (1, 2, 4, 5)$

networks, web graphs, bioinformatics, and graph databases.

Graph Partitions

Many problems in graph analytics can be sped up if graphs can be partitioned. A *k-partition* takes a graph $G(V, E)$ and divides vertex set V into k disjoint subsets V_i such that.

$$\bigcup_{i=1}^k V_i = V. \tag{7}$$

The usefulness of a partition is often judged positively by its evenness or size balance and negatively by the number of edges that are cut. Edge cuts result when an edge ends up crossing from one vertex subset to another. Each part of a partitioned graph is stored in a separate graph (either on a server with a large memory or to multiple servers in a cluster). Algorithms can then work in parallel on smaller graphs and combine results to solve the original problem. The extra work required to combine results is related to the number of cuts done in partitioning.

Although finding balanced min-cut partitions is an NP-hard problem, there are practical

algorithms and implementations that do an effective job on very large graphs. One of the better software packages for graph partitioning is METIS (Karypis and Kumar 1995) as it tends to provide good balance with fewer edge cuts than alternative software. “METIS works in three steps: (1) coarsening the graph, (2) partitioning the coarsened graph, and (3) uncoarsening” (Wang et al. 2014). Faster algorithms that often result in more edge cuts than METIS include random partitioning and ordered partitioning, while label propagation partitioning trades off fewer edge cuts for less balance.

Related topics in graph analytics include graph clustering and finding graph components. In *graph clustering*, vertices that are relatively more highly interconnected are placed in the same cluster, e.g., friend groups. A subgraph is formed by including all edges (u, v) for which u and v are in the same cluster. At the extreme end, vertices could be grouped together, so long as there exists a *path* (u, v) between any two vertices u and v in the group. The subgraphs formed from these groups are referred to as *strongly connected components*. Further, the subgraphs are referred to as *weakly connected components*, if there is a path between any two vertices in the group in the underlying undirected graph (where directionality of edges is ignored).

Graph Databases

A very large vertex and edge labeled multi-digraph where the labels are rich with information content can be viewed as a graph database. Property graphs, often mentioned in the literature, are extensions where a label (or property) is allowed to have multiple attributes (e.g., name, address, phone). For a graph database, the following capabilities should be provided: (1) persistent storage (should be able to access without complete loading of a file), (2) update/transactional capability, and (3) high-level query language. When data can be organized in the form of a graph, query processing in a graph database can be much faster than the alternative of converting the graph into relations stored in a relational database.

Examples of graph databases include Neo4j, OrientDB, and Titan (Angles 2012). In addition, Resource Description Framework (RDF) stores used in the Semantic Web are very similar to graph databases (certain restricted forms would qualify as graph databases). Query languages include Cypher, Gremlin, and SPARQL. The following example query written in the Cypher language (used by Neo4j) expresses the graph pattern discussed in the section “[Introduction](#)”:

MATCH (u: PhDCandidate, v: PostDoc, w: Professor,

u – [:FOLLOWS] –> w,
v – [:CITES] –> w,
w – [:READS] –> v)

The answer would be all (or a sufficient number of) matching patterns found in the graph database, with vertex variables, u , v , and w , replaced by actual ResearchGate users.

Query processing and optimization for graph databases (Gubichev 2015) include parsing a given query expressed in the query language, building an evaluation-oriented abstract syntax tree (AST), optimizing the AST, and evaluating the optimized AST. Bottom-up evaluation could be done by applying algorithms for graph algebra operators (e.g., selection, join, and expand) (Gubichev 2015). Where applicable, pattern matching algorithms discussed in section “[Graph](#)

[Patterns](#)” may be applied as well. In Neo4j, query processing corresponds to the subgraph isomorphism problem, while for RDF/SPARQL stores, it corresponds to the homomorphism problem (Gubichev 2015).

Conclusions

Graph analytics and databases are growing areas of interest. A brief overview of these areas has been given. More detailed surveys and historical background may be found in the following literature: A historical view of graph pattern matching covering exact and inexact pattern matching is given in (Conte et al. 2004). Research issues in big data graph analytics is given in (Miller et al. 2015). A survey of big data frameworks supporting graph analytics, including Pregel and Apache Giraph, is given in (Batarfi et al. 2015).

Further Reading

- Angles, R. (2012). A comparison of current graph database models. In *IEEE 28th international conference on data engineering workshops (ICDEW)* (pp. 171–177). Washington, DC: IEEE.
- Batarfi, O., ElShawi, R., Fayoumi, A., Nouri, R., Beheshti, S. M. R., Barnawi, A., & Sakr, S. (2015). Large scale graph processing systems: Survey and an experimental evaluation. *Cluster Computing*, 18(3), 1189–1213.
- Conte, D., Foggia, P., Sansone, C., & Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03), 265–298.
- Fard, A., Nisar, M. U., Ramaswamy, L., Miller, J. A., & Saltz, M. (2013). A distributed vertex-centric approach for pattern matching in massive graphs. In *IEEE international conference on Big Data* (pp. 403–411). Washington, DC: IEEE.
- Fard, A., Manda, S., Ramaswamy, L., & Miller, J. A. (2014a). Effective caching techniques for accelerating pattern matching queries. In *IEEE international conference on Big Data (Big Data)* (pp. 491–499). Washington, DC: IEEE.
- Fard, A., Nisar, M. U., Miller, J. A., & Ramaswamy, L. (2014b). Distributed and scalable graph pattern matching: Models and algorithms. *International Journal of Big Data*, 1(1), 1–14.
- Gubichev, A. (2015). *Query processing and optimization in graph databases* (PhD thesis). Technische Universität München, München.

- Henzinger, M. R., Henzinger, T. A., & Kopke, P. W. (1995). Computing simulations on finite and infinite graphs. In *Proceedings, 36th annual symposium on foundations of computer science* (pp. 453–462). Washington, DC: IEEE.
- Karypis, G., & Kumar, V. (1995). Analysis of multilevel graph partitioning. In *Proceedings of the 1995 ACM/IEEE conference on supercomputing* (p. 29). New York: ACM.
- Ma, S., Cao, Y., Fan, W., Huai, J., & Wo, T. (2011). Capturing topology in graph pattern matching. *Proceedings of the VLDB Endowment*, 5(4), 310–321.
- Ma, S., Cao, Y., Fan, W., Huai, J., & Wo, T. (2014). Strong simulation: Capturing topology in graph pattern matching. *ACM Transactions on Database Systems (TODS)*, 39(1), 4.
- Miller, J. A., Ramaswamy, L., Kochut, K. J., & Fard, A. (2015). Directions for big data graph analytics research. *International Journal of Big Data (IJBD)*, 2(1), 15–27.
- Tsourakakis, C. E., Kang, U., Miller, G. L., & Faloutsos, C. (2009). Doulion: counting triangles in massive graphs with a coin. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 837–846). New York: ACM.
- Wang, L., Xiao, Y., Shao, B., & Wang, H. (2014). How to partition a billion-node graph. In *IEEE 30th International Conference on data engineering (ICDE)* (pp. 568–579). Washington, DC: IEEE.

Harnessing the Data Revolution

► [Big Data Research and Development Initiative \(Federal, U.S.\)](#)

HDR

► [Big Data Research and Development Initiative \(Federal, U.S.\)](#)

Health Care Delivery

Paula K. Baldwin
Department of Communication Studies, Western
Oregon University, Monmouth, OR, USA

Peter Groves, Basel Kayyali, David Knott, and Steve Van Kuiken note that the evolution and use of big data are in its formative stages, and its true potential has yet to be revealed. Mike Cottle, Waco Hoover, Shadaab Kanwal, Marty Kohn, Trevor Strome, and Neil W. Treister write that in 2011, US health care data totaled 150 exabytes, and that number is increasing. To put that figure in

a relatable perspective, five exabytes equals 10^{18} gigabytes and that is calculated to be the sum of all words in the human vocabulary. In addition, Cottle, Hoover, Kanwal, Kohn, Strome, and Treister note that there are five separate categories of big data relating specifically to health and health care delivery. First, there are web and social media data that also include health plan websites and smart phone apps, to name a few. Second, there are the machine-to-machine data that originate from sensors, meters, and other devices. Third on the list is big transaction data consisting of health care claims and other billing records. Fourth is biometric data consisting of fingerprints, genetics, handwriting, retinal scans, and other similar types of data, including x-rays and other types of medical imaging. Finally, there is the data generated by electronic medical records (EMRs), health care providers' notes, electronic correspondence, and paper documents.

Other industries such as retail and banking embraced utilizing big data to benefit both the organization and the consumer, but the health care industry is behind in that process. Catherine M. DesRoches, Dustin Charles, Michael F. Furukawa, Maulik S. Joshi, Peter Kralovec, Farzad Mostashari, Chantal Worzala, and Ashish K. Jha report that in 2012, only 44% of US hospitals reported using a basic electronic health records (EHRs) system and rural and nonteaching

hospitals lag significantly behind in adopting EHRs systems.

As the exploration of possible uses of big data in health care delivery continues, the potential increases to affect both the health care provider and the health care recipient, positively. However, as Groves, Kayyali, Knott, and Van Kuiken write, the health care industries suffer from several inhibitors: resistance to change, lack of sufficient investment in technology, privacy concerns for health care recipients, and lack of technology for integrating data across multiple systems.

Creating Health Care Delivery Systems

IBM's Information Management identified three important areas to a successful health care delivery transition. First, build health care systems that can efficiently manage resources and improve patient care while reducing the cost of care. Second, health care organization should focus on improving quality and efficiency care by focusing on a deep understanding of health care recipients' needs. Finally, in order to fully engage with all segments of the US population, emphasis on increasing access to health care is crucial.

Health Care Delivery System Benefits

The Healthcare Leadership Council identified three key benefits for health care recipients.

First, both individuals and families will have multiple options available to them for their health care delivery with the focus being on the "right treatment at the right time in the right place to each patient." Second, shifting the emphasis of health care delivery to better long-term values rather than the current strategy of reduction of short-term costs will provide more economic relief for health care recipients. Finally, the development and implementation of a universal health care delivery system will create a major shift in the health care industry itself, away from an industry made up of disparate parts and toward an integrated model for

providing premium care for all health care recipients.

Challenges for Health Care Delivery Systems

The Healthcare Leadership Council identified two areas critical to the future use and implementation of big data into health care delivery systems. First, a new platform for linking health care recipients' medical records and health care must be developed, and second, the USA must make a serious commitment in health care research and development as well as the education of future generations of health care providers. Erin McCann seconds that and writes that the biggest challenge in effectively caring for patients today is data on patients comes from different institutions and different states all using multiple data tracking systems. In order for the patient information to be useful, technology must develop a universal platform through which the various tracking systems and electronic health records (EHRs) can communicate accurately.

Communication between the technology and the health care provider is challenged and driven by the health care recipients themselves as they change doctors, institutions, or insurance. These changes are driven by changes in locale, changes in health care needs, and other life changes; therefore, patient engagement in the design of these changes is paramount. In order for health care delivery to be successful, the communication platform must be able to identify and adapt to those changes. The design and implementation of a common platform for the different streams of medical information continue to evolve as the technology advances. With these advances, health care recipients in rural and urban settings will have equal access to health care.

Cross-References

- ▶ [Electronic Health Records \(EHR\)](#)
- ▶ [Health Care Delivery](#)
- ▶ [Health Informatics](#)

Further Reading

- Cottle, M., et al. (2013). *Transforming health care through big data: Strategies for leveraging big data in the health care industry*. Institute for Health Technology Transformation. Washington, D.C.
- DesRoches, C. M., et al. (2013). Adoption of electronic health records grows rapidly, but fewer than half of U.S. hospitals had at least a basic system. *Health Affairs*, 32(8), 1478–1485.
- Groves, P., et al. (2014). The ‘Big Data’ revolution in healthcare: Accelerating value and innovation. Center for US health system reform business technology office, McKinsey & Company.
- Healthcare Leadership Council. Key Issues. <http://www.hlc.org/key-issues/> (n.d.). Accessed Nov 2014.
- IBM. Harness your data resources in healthcare. Big data at the speed of business. <http://www-01.ibm.com/software/data/bigdata/industry-healthcare.html> (n.d.). Accessed Nov 2014.
- McCann, E. (2014). No interoperability? Goodbye big data. Healthcare IT news.
- McCarthy, R. L., et al. (2012). *Introduction to health care delivery*. Sudbury/Mass: Jones & Bartlett Learning.

Health Informatics

Erik W. Kuiler
George Mason University, Arlington, VA, USA

Background

The growth of informatics as a technical discipline reflects the increased computerization of business operations in both the private and government sectors. Informatics focus on how information technologies (IT) are applied in social, cultural, organizational, and economic settings. Although informatics have their genesis in the mainframe computer era, it has only been since the 1980s that health informatics, as Marsden S. Blois notes, have gained recognition as a technical discipline by concentrating on the information requirements of patients, health-care providers, and payers. Health informatics also support the requirements of researchers, vendors, and oversight agencies at the federal, state, and local levels.

The health informatics’ domain is extensive, encompassing not only patient health and continuity of care but also epidemiology and public health. Due to the increased use of IT in health-care delivery and management, the purview of health informatics is expected to continue to grow.

The advent of the Internet; the availability of inexpensive high-speed computers, voice recognition and mobile technologies, and large data sets in diverse formats from diverse sources; and the use of social media have provided opportunities for health-care professionals to incorporate IT applications in their practices. In the United States, the impetus for health informatics came during 2008–2010. Under the Health Information Technology (HIT) for Economic and Clinical Health (HITECH) component of the American Recovery and Reinvestment Act of 2009 (ARRA), the Centers for Medicare and Medicaid Services (CMS) reimburse health service providers for using electronic documents in formats certified to comply with HITECH’s Meaningful Use (MU) standards. The Patient Protection and Affordable Care Act of 2010 (ACA) promotes access to health care and greater use of electronically transmitted documentation. Health informatics are expected to provide a framework for the electronic exchange of health information that complies with all legal requirements and standards.

The increased acceptance of HIT is also expected to expand the delivery of comparative effectiveness- and evidence-based medicine. However, while there are important benefits to health data sharing among clinicians, caregivers, and payers, the rates of HIT technical advances have proven to be greater than their rates of assimilation.

Electronic Health Records

Electronic health documentation overcomes the limitations imposed by paper records: idiosyncratic interpretability, inconsistent formats, and indifferent quality of information. Electronic

health documentation usually takes one of three forms, each of which must comply with predetermined standards before they are authorized for use: electronic medical records (EMR), electronic health records (EHRs), and personal health records (PHR). The National Alliance for Health Information Technology distinguishes them as follows (2008): An EMR provides information about an individual for use by authorized personnel within a health-care organization. A PHR provides health-care information about an individual from diverse sources (clinicians, caregivers, insurance providers, and support groups) for the individual's personal use. An EHR provides health-related information about an individual that may be created, managed, and exchanged by authorized clinical personnel. EHRs may contain both structured and unstructured data so that it is possible to share coded diagnostic data, clinician's notes, personal genomic data, and X-ray images in the same document, with substantially less likelihood of error in interpretation or legibility.

Health Level 7 (HL7), an international organization, has promulgated a set of document architectural standards that enable the creation of consistent electronic health documents. The Consolidated Clinical Document Architecture (C-CDA) and the Quality Reporting Document Architecture (QRDA) provide templates that reflect the HL7 Reference Information Model (RIM) and can be used to structure electronic health documents. The C-CDA consolidates the initial CDA with the Continuity of Care Document (CCD) developed by the Healthcare Information Technology Standards Panel (HITSP). The C-CDA and the QRDA support the Extensible Markup Language (XML) standard so that any documents developed according to these standards are both human- and machine-readable. C-CDA and QRDA documents may contain structured and unstructured data. HL7 recommends the use of data standards, such as the Logical Observation Identifiers Names and Codes (LOINC), managed by the Regenstrief Institute, to ensure consistent interpretability.

Health Information Exchange

With electronic health records, health information exchange (HIE) provides the foundation of health informatics. Adhering to national standards, HIE operationalizes the HITECH MU provisions by enabling the electronic conveyance of health information among health-care organizations. Examples are case management and referral data, clinical results (laboratory, pathology, medication, allergy, and immunization data), clinical summaries (CCD and PHR extracts), images (including radiology reports and scanned documents), free-form text (office notes, discharge notes, emergency room notes), financial data (claims and payments), performance metrics (providers and institutions), and public health data.

The US Department of Health and Human Services Office of the National Coordinator (DHHS ONC) has established the eHealth Exchange as a network of networks to support HIE by formulating the policies, services, and standards that apply to HIE. HL7 has produced the Fast Health Interoperable Resources (FHIR) framework for developing web-based C-CDA and QRDA implementations that comply with web standards, such as XML, JSON, and HTTP. An older standard, the American National Standards Institute X12 Electronic Data Interchange (ANSI X12 EDI), supports the transmission of Health Care Claim and Claim Payment/Advice data (transactions 835 and 837).

Health Domain Data Standards

To ensure semantic consistency and data quality, the effective use of health informatics depends on the adoption of data standards, such as the internationally recognized Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), maintained by the International Health Terminology Standards Development Organisation (IHTSDO), a multilingual lexicon that provides coded clinical terminology extensively used in EHR management. RxNorm, maintained by the National Institutes of Health's National Library of Medicine (NIH NLM), provides a common

(“normalized”) nomenclature for clinical drugs with links to their equivalents in other drug vocabularies commonly used in pharmacology and drug interaction research. The Logical Observation Identifiers Names and Codes (LOINC), managed by the Regenstrief Institute, provides a standardized lexicon for reporting lab results. The International Classification of Diseases, ninth and tenth editions, (ICD-9 and ICD-10), are also widely used.

Health Data Analytics

The accelerated adoption of EHRs has increased the availability of health-care data. The availability of large data sets, in diverse formats from different sources, is the norm rather than the exception. By themselves data are not particularly valuable unless researchers and analysts can discern patterns of meaning that collectively constitute information useful to meet strategic and operational requirements. Health data analytics, comprising statistics-based descriptive and predictive modeling, data mining, and text mining, supported by natural language processing (NLP), provide the information necessary to improve population well-being. Data analytics can help reduce operational costs and increase operational efficiency by providing information needed to plan and allocate resources where they may be used most effectively. From a policy perspective, data analytics are helpful in assessing programmatic successes and failures, enabling the modification and refinement of policies to effect their desired outcomes at an optimum level.

Big Data and Health Informatics

Big Data, with their size, complexity, and velocity, can be beneficial to health informatics by expanding, for example, the range and scope of research opportunities. However, the increased availability of Big Data has also increased the need for effective privacy and security management, defense against data breaches, and data storage management. Big Data retrieval and

ingestion capabilities have increased the dangers of unauthorized in-transit data extractions, transformations, and assimilation unbeknownst to the authorized data owners, stewards, or recipients. To ensure the privacy of individually identifiable health information, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) requires health data records to be “anonymized” by removing all personally identifiable information (PII) prior to their use in data analytics. As data analytics and data management tools become more sophisticated and robust, the availability of Big Data sets will increase. Issues affecting the management and ethical, disciplined use of Big Data will continue to inform policy discussions. With the appropriate safeguards, Big Data analytics enhance our capabilities to capture program performance metrics focused on costs, comparative effectiveness of diagnostics and interventions, fraud, waste, and abuse.

Challenges and Future Trends

Health informatics hold the promise of improving health care in terms of access and outcomes. But many challenges remain. For example, Big Data analytics tools are in their infancy. The processes to assure interorganizational data quality standards are not fully defined. Likewise, anonymization algorithms need additional refinement to ensure the privacy and security of personally identifiable information (PII). In spite of the work that still needs to be done, the importance of health informatics will increase as these issues are addressed, not only as technical challenges but also to increase the social good.

Further Reading

- Falick, D. (2014). For big data, big questions remain. *Health Affairs*, 33(7), 1111–1114.
- Miller, R. H., & Sim, I. (2004). Physicians’ use of electronic medical records: Barriers and solutions. *Health Affairs*, 23(2), 116–126.
- Office of the National Coordinator. (2008). *The National Alliance for Health Information Technology Report to the National Coordinator for Health Information*

Technology on Defining Key Health Information Technology Terms. Health Information Technology. Available from <http://www.hitechanswers.net/wp-content/uploads/2013/05/NAHIT-Definitions2008.pdf>.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2(3), 1–10. Available from <http://www.hissjournal.com/content/2/1/3>.

Richesson, R. L., & Krischer, J. (2007). Data standards in clinical research: Gaps, overlaps, challenges and future directions. *Journal of the American Medical Informatics Association*, 14(6), 687–696.

High Dimensional Data

Laurie A. Schintler

George Mason University, Fairfax, VA, USA

Overview

While big data is typically characterized as having a massive number of observations, it also refers to data with high dimensionality. High-dimensional data contains many attributes (variables) relative to the sample size, including instances where the number of attributes exceeds the number of observations. Such data are common within and across multiple domains and disciplines, from genomics to finance and economics to astronomy. Some examples include:

- Electronic Health Records, where each record contains various data points about a patient, including demographics, vital signs, medical history, diagnoses, medications, immunizations, allergies, radiology images, lab and test results, and other items
- Earth Observation Data, which contains locational and temporal measurements of different aspects of our planet, e.g., temperature, rainfall, altitude, soil type, humidity, terrain, etc.
- High-Frequency Trading data, which comprises real-time information on financial transactions and stock prices, along with unstructured content, such as news and social

media posts, reflecting consumer, and business sentiments, among other things

- Micro-array data, where each microarray comprises tens of thousands of genes/features, but there are only a limited number of clinical samples
- Unstructured documents, where each document contains numerous words, terms, and other attributes

High dimensional data raise unique analytical, statistical, and computational issues and challenges. Data with both a high number of dimensions and observations raises an additional set of issues, particularly in terms of algorithmic stability and computational efficiency. Accordingly, the use of high-dimensional data requires specific kinds of methods, tools, and techniques.

Issues and Challenges

Regression models based on high dimensional data are vulnerable to statistical problems, including noise accumulation, spurious correlations, and incidental endogeneity. Noise can propagate in models with many variables, particularly if there is a large share of poor predictors. Additionally, uncorrelated random variables in such models also can show a strong association in the sample, i.e., there is the possibility for spurious correlations. Finally, in large multivariate models, covariates may be fortuitously correlated with the residuals, which is the essence of incidental endogeneity. These issues can compromise the validity, reliability, interpretability, and appropriateness of regression models. In the particular case where the number of attributes exceeds the sample size, there is no longer a unique least-squares solution, as the variance of each of the estimators becomes infinite.

Another related problem is the “curse of dimensionality,” which has implications for the accuracy and generalizability of statistical learning models. For supervised machine learning, a model’s predictive performance hinges critically on how well the data used for training accurately reflects the phenomenon being modeling. In this

regard, the sample data should contain a representative combination of predictors and outcomes. However, high-dimensional data tends to be sparse, a situation in which the training examples given to the model fail to capture all possible combinations of the predictors and outcomes, including infrequent occurrences. This situation can lead to “overfitting,” where the trained model has poor predictive performance when using data outside the training set. As a general rule, the amount of data needed for accurate model generalization increases exponentially with the dimensionality.

In instances where high-dimensional data contains a large of observations, model optimization can be computationally expensive. Accordingly, when scalability and computational complexity must be considered in selecting models when working with such data.

Strategies and Solutions

Two approaches for addressing the problems associated with high-dimensional data involve reducing the dimensionality of the data before being analyzed or selecting models specifically designed to handle high dimensional data.

Subset or Feature Selection

This strategy involves the removal of irrelevant or redundant variables from the data prior to modeling. As feature selection keeps only a subset of original features, it has the advantages of making the final model more interpretable and minimizing costs associated with data processing and storage. Feature selection can be accomplished in a couple of different ways, each of which has advantages and disadvantages. One tactic is to simply extract predictors that we believe are most strongly associated with the output. However, the drawback of this technique is that it requires a priori knowledge on what are appropriate predictors, which can be difficult when working with massive numbers of variables. An alternative is to apply the “best subset” selection, which fits separate regression models for each possible combination of predictors. In this approach, we fit all possible models

that contain precisely one predictor, then move on to models with exactly two predictors, and so on. We then examine the entire collection of models to see which one performs best while minimizing the number of covariates in the model. Indeed, this is a simple and intuitively appealing approach. However, this technique can become computationally intractable when there are large numbers of variables. Further, the larger the search space, the higher the chance of finding models that look good on the training data but have low predictive power. An enormous search space can lead to overfitting and high variance of the coefficient estimates. For these reasons, stepwise methods – forward or backward elimination – are attractive alternatives to best subset selection.

Shrinkage Methods

Shrinkage (regularization) methods involve fitting a model with all the predictors, allowing for small or even null values for some of the coefficients. Such approaches not only help in selecting predictors but also reduce model variance, in turn reducing the chances of overfitting. Ridge regression and lasso regression are two types of models, which utilize regularization methods to “shrink” the coefficients. They accomplish this through the use of a penalty function. Ridge regression includes a weight in the objective function used for model optimization to create a penalty for adding more predictors. This has the effect of shrinking one or more of the coefficients to values close to zero. On the other hand, lasso regression uses the absolute values of the coefficients in the penalty function, which allows for coefficients to go to zero.

Dimensionality Reduction

Another approach for managing high-dimensional data is to reduce the complexity of the data prior to modeling. With dimensionality reduction methods, we do not lose any of the original variables. Instead, all the variables get folded into the high-order dimensions extracted. There are two categories of dimensionality reduction techniques. In data-oblivious approaches, we do the dimensionality-reducing mapping without using the data or knowledge about the data.

Random projection and sketching are two popular methods in this category. The advantages of the data-oblivious approach are that (1) it is not computationally intensive and (2) it does not require us to “see” or understand the underlying data. Data-aware reduction maps the data without explicit consideration of its contents; instead, it “learns” the data structure. Principal Component Analysis (PCA) and clustering algorithms are examples of such dimensionality reduction methods.

Other Methods

Support Vector Machines (SVM), a machine learning algorithm, is also suitable for modeling high-dimensional data. SVMs are not sensitive to the data’s dimensionality, and they can effectively deal with noise and nonlinearities. Ensemble data analysis can reduce the likelihood of overfitting models based on high-dimensional data. Such methods use multiple, integrated algorithms to extract information from the entire data set. Bootstrapping, boosting, bagging, stacking, and random forests are all involved in ensemble analysis.

Concluding Remarks

While high-dimensional big data provide rich opportunities for understanding and modeling complex phenomena, its use comes with various issues and challenges, as highlighted. Certain types of high dimensional big data – e.g., spatial or network data – can contribute to additional problems unique to the data’s particular nuances. Accordingly, when working with high-dimensional data, it is imperative first to understand and anticipate the specific issues that may arise in modeling the data, which can help optimize the selection of appropriate methods and models.

Cross-References

- [Data Reduction](#)
- [Ensemble Methods](#)

Further Reading

- Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. New York: Springer.
- Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1), 101.
- Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2), 293–314.
- Genender-Feltheimer, A. (2018). Visualizing high dimensional and big data. *Procedia Computer Science*, 140, 112–121.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: Springer.

HIPAA

William Pewen

Department of Health, Nursing and Nutrition,
University of the District of Columbia,
Washington, DC, USA

As health concerns are universal, and health care spending is escalating globally, big data applications offer the potential to improve health outcomes and reduce expenditures. In the USA, access to an individual’s health data including both clinical and fiscal records has been regulated under the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The Act established a complex and controversial means of regulating health information which has been both burdensome to the health sector and has failed to fully meet public expectations for ensuring the privacy and security of information.

In the absence of a broad statutory regime addressing privacy, legislative efforts to address the public concern regarding health information have been sector specific and reactive. Standards have relied upon a foundation of medical ethics beginning with the Hippocratic Oath. Yet even such a recognized standard has not been a static one, as modern versions of the oath exhibit substantial changes from the original form. The development of federal patient protections drew

substantially from reforms established in both the Nuremberg Code and the Helsinki Declaration, with the latter undergoing periodic revisions, including recent recognition that the disclosure of “identifiable human material or identifiable data” can pose substantial risks to individuals.

Enactment

Enactment of HIPAA in 1996 provided a schema for the regulation of the handling of individually identifiable health information by “covered entities” – those medical service providers, health plans, and certain other organizations involved in treatment and related financial transactions. Such information includes critical primary data on a healthcare sector now impacting over 325 million Americans and involving over \$3 trillion in annual spending.

HIPAA regulates the disclosure and use of health data, rather than its collection, and functions primarily as a tool for maintaining confidentiality. The pursuance of rules to implement the Act spanned six contentious years ending in 2002 after requirements for active consent by individuals prior to data disclosures were substantially reduced.

A critical construct of HIPAA is the concept of protected health information (“PHI”), which is defined as “individually identifiable health information.” Only PHI is protected under the Act. HIPAA provides for required disclosure of PHI when requested by the patient or by the Secretary of Health and Human Services for certain defined purposes including audit and enforcement. Aside from this, disclosure is permitted for the purposes of treatment, payment, and healthcare operations without the need for specific patient authorization.

HIPAA also provides for significant exceptions under which consent for individual data disclosure is not required. Notable among these is a public health exception under which the use of data without consent for reporting and research is permitted. In addition, after implementation of the 1996 Act, many diverse activities were conducted as “healthcare operations,” ranging from a covered entity’s quality assurance efforts to unsolicited marketing to patients.

Under HIPAA personal health information may be “de-identified” by removal of 18 specified identifiers or by a process in which expert certification is obtained to ensure a low probability of identification of an individual patient. Such de-identified data is no longer considered PHI and is not protected under HIPAA.

HITECH Amendments to HIPAA

The enactment of the Health Information Technology for Economic and Clinical Health (HITECH) Act resulted in a number of changes to HIPAA. These focused primarily on bringing the business associates of covered entities under HIPAA regulation, increasing penalties for violations, establishing notification and penalties for data breaches, and limiting the sale of PHI without a patient’s consent. However, three key aspects remained essentially intact.

First, the public health exception remained and explicitly permits the sale of PHI for research purposes – now including sale by covered entities for public health purposes at a profit. This provides a major potential data stream for big data applications development. Second, the de-identification scheme was preserved as a construct to utilize data outside the limitations imposed for PHI (although some guidance in the use of de-identified and limited data sets has been issued). Finally, the scope of HIPAA remains limited to health data handled by covered entities. A vast spectrum of data sharing including health-related websites, organizations, and even “personal health records” remain outside HIPAA regulation.

Big Data Utilization of HIPAA Data

The healthcare applications for which big data offers the most immediate and clear promise are those aimed at utilizing data to realize improved outcomes and cost savings. Research and innovation to achieve such advances rely on a range of data sources but in particular involve access to the electronic health record (EHR) and claims and payment data. Both lie within the scope of

HIPAA regulation. Given that access to primary data is of critical importance for big data healthcare applications, three strategies for access under the HIPAA regime are evident.

The first of these utilizes a provision of the HITECH Act which addressed the “healthcare operations” exception and provides means of conducting internal quality assurance and process improvement within a covered entity. This facilitates a host of big data applications which may be employed to analyze and improve care. However, this strategy presents issues of both ethical consent and legal access in any subsequent synthesis of data involving multiple covered entities.

A second strategy involves the use of de-identified data. HIPAA’s requirements for the utilization of such data are minimal, and use does not require the consent of individual patients. However, this strategy presents two major problems linked to the means of de-identification of PHR. The first of these is that, for many studies, the loss of the 18 designated identifiers may compromise the aims of the project. Identifying fields can be critical to an application. Attempts at longitudinal study are particularly impacted by such a strategy. Alternatively study may either discard the designated identifiers or rely upon the use of a certifying party that the data has been de-identified to the extent that re-identification is highly unlikely – yet the latter approach may offer an uncertain shield with regard to ultimate liability.

The use of de-identified data also presents a clear contradiction when relied upon for big data applications. The contradiction must be noted that one key aspect of big data involves the ability to link disparate data sets; consequently, this key attribute undermines the fundamental premise of de-identification. To the extent that big data techniques are effective, the “safe harbor” of de-identification is thus negated.

A third strategy is indisputably HIPAA compliant. The acquisition of consent from individuals does not rely upon uncertain de-identification nor is it constrained to data sets contained only within a single covered entity. While consent mechanisms undoubtedly could be constructed as to be more concise, and the process less

cumbersome, the use of consent also addresses a key observation underlying continued debate around HIPAA: the finding of the Institute of Medicine that only approximately 10 percent of Americans support access to health data for research without their consent.

Data Outside the Scope of HIPAA

The original drafting of HIPAA occurred in a context in which electronic health data exchange was in its infancy. The HITECH Act remains constrained by the HIPAA construct and consequently does not address health data outside of the “covered entity” construct. HIPAA thus fails to regulate collection, disclosure, or the use of data if there is no covered entity relationship. While access to the most desirable data may be protected by HIPAA, a huge expanse of health-related information is not, such as purchases of over-the-counter drugs and data sharing on most health-related websites.

Nonregulated data is highly variable in validity, reliability and precision - raising concerns regarding its application in the study of health states, outcomes, and costs. Such data may be relatively definitive, such as information shared by the patient and information gleaned through commercial transactions such as consumer purchases. A report that Target Stores developed a highly accurate means of identifying pregnant women through retail purchase patterns is illustrative of the power of secondary data in deriving health information which would otherwise be protected under HIPAA.

The use of such surrogate data presents a host of problems, including both public objection to the use of technology to undermine the statutory intent of HIPAA, as well as the application of big data in facilitating discrimination which could evade civil rights protections. In a context in which the majority of Americans lack confidence in the HIPAA framework for maintaining the confidentiality and security of individual health information, the current regulatory framework may not remain static.

Cross-References

- Biomedical Data
- De-identification/Re-identification
- Health Informatics
- Patient Records

Further Reading

- Caines, K., & Hanania, R. (2013). Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association*, 20, 7–15.
- Duhigg, C. (2012). *How companies learn your secrets*. New York Times. 16 Feb 2012.
- Pewen W. *Protecting our civil rights in the era of digital health*. The Atlantic. 2 Aug 2012. <http://www.theatlantic.com/health/archive/2012/08/protecting-our-civil-rights-in-the-era-of-digital-health/260343/>. Accessed Aug 2016.
- 110 Stat. 1936 – Health Coverage Availability and Affordability Act of 1996. <https://www.gpo.gov/fdsys/granule/STATUTE-110/STATUTE-110-Pg1936>. Accessed Sept 2017
- U.S. Department of Health And Human Services. *Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule*. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-identification/guidance.html>. Accessed Sept 2017.
- World Medical Association. *Declaration of Helsinki – Ethical principles for medical research involving human subjects* <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>. Accessed Sept 2017.

Human Resources

Lisa M. Frehill
Energetics Technology Center, Indian Head,
MD, USA

Human resources (HR) management is engaged and applied throughout the full employee lifecycle, including recruitment and hiring, talent management and advancement, and exit/retirement. HR includes operational processes and,

with the expansion of the volume, variety, and velocity of data in the past 20 years, more emphasis has been placed on strategic planning and future-casting. As an organizational function that has long embraced the use of information technology and data, HR is well-positioned to deploy big data in many ways, but also faces some challenges.

Enterprise data warehouses (EDW) have been a key tool for HR for many years, with the more recent development of interactive HR dashboards to enable managers to access employee data and analytics (business intelligence, or BI tools) to monitor key metrics such as turnover, employee engagement, workforce diversity, absenteeism, productivity, and the efficiency of hiring processes among others. In the past decade, the availability of big data has meant that organizational data systems now ingest unstructured, text-based, and naturally occurring data from our increasingly digital world. While EDW were designed to function well with structured, fixed semantics data, much of big data, with highly variable semantics, necessitates harmonization and analysis in “data lakes” prior to being operationally useful to organizations.

There are a number of ways HR uses big data. Examples include, but are not limited to, the following:

- *Recruitment and hiring*: Organizations use big data to manage their brands to potential employees, making extensive use of social media platforms (e.g., Twitter, LinkedIn, and Facebook). These platforms are also sources of additional information about potential employees, with many organizations developing methods of ingesting data from applicants’ digital presence and using these data to supplement the information available from résumés and interviews. Finally, big data has provided tools for organizations to increase the diversity of those they recruit and hire by both providing a wider net to increase the size and variety of applicant pools and to counter interviewers’ biases by gathering and analyzing observational data in digital interviews.

- *Talent management*: There are many aspects of this broad HR function, which involves performance and productivity management, workforce engagement, professional development and advancement, and making sure the salary and benefits portfolio keeps pace with the rewards seen as valuable by the workforce, among others. For example, big data has been cited as a valuable tool for identifying skills gaps within an organization's workforce in order to determine training that may need to be offered. Some organizations have employed gamification strategies to gather data about employee skills. Finally, big data provides the means for management to exert control over the workforce, especially a geographically dispersed one, such as is common within the coronavirus pandemic of 2020–2021.
- *Knowledge management*: Curation of the vast store of organizational information is another critical HR big data task. Early in the life of the World Wide Web, voluminous employee handbooks moved online even as they became larger, more detailed, and connected with online digital forms associated with a multitude of organizational processes (e.g., wage and benefits, resource requisitions, performance management). Big data techniques have been important in expanding these knowledge management functions. They have also expanded on these traditional functions to include capturing and preserving institutional knowledge for more robust succession planning, which has recently been important as baby boomers enter retirement.

Big data poses important challenges for HR professionals. First, issues of privacy and transparency are important as more data about people are more rapidly available. For example, in the past several years, employees' behavior outside the workplace has been more easily surveilled by employers via the proliferation of social media platforms, with consequential outcomes for employees whose behavior is considered inappropriate by their employers. Additionally, algorithms used with ingested digital data and problems associated with different ways various

devices process application information (including assessment tests of applicants) run a risk of reducing the transparency of hiring processes and of inadvertently introducing the biases HR professionals hope to reduce.

Second, the proliferation of big data, the transition from EDW to data lakes, and the greater societal pressure on organizations to be more transparent with respect to human resources means HR professionals need additional data analytics skills. HR professionals need to understand limitations associated with big data such as quality issues (e.g., reliability, validity, and bias), but they also need to be able to more meaningfully connect data with organizational outcomes without falling into the trap of spurious results.

In closing, big data has been an important resource for data-intensive HR organizational functions. Such unstructured, natural data has provided complementary information to the highly structured and designed data HR professionals have long used to recruit, retain, and advance employees necessary for efficient and productive organizations.

Disclaimer The views expressed in this entry are those of the author and do not necessarily represent the views of Energetics Technology Center, the Institute of Museum and Library Services, the U.S. Department of Energy, or the government of the United States.

Bibliography/Further Readings

- Corritore, M., Goldberg, A., & Srivastava, S. B. (2020, January). The new analytics of workplace culture. SHRM online at: <https://shrm.org/resourcesandtools/hr-topics/technology/pages/the-new-analytics-of-workplace-culture.aspx>.
- Friedman, T., & Heudecker, N. (2020, February). Data hubs, data lakes and data warehouses: How they are different and why they are better together. Gartner online at: <https://www.gartner.com/doc/reprints?id=1-24IZJZ2F&ct=201103&st=sb>.
- Garcia-Arroyo, J., & Osca, A. (2019). Big data contributions to human resource management: A systematic review. *International Journal of Human Resource Management*. <https://doi.org/10.1080/09585192.2019.1674357>.
- Giacumo, L. A., & Breman, J. (2016). Emerging evidence on the use of big data and analytics in workplace

- learning: A systematic literature review. *The Quarterly Review of Distance Education*, 17(4), 21–38.
- Howard, N., & Wise, S. (2018). *Best practices in linking data to organizational outcomes*. Bowling Green: Society for Industrial and Organizational Psychology (SIOP). Online at: <https://www.siop.org/Portals/84/docs/White%20Papers/Visibility/DataLinking.pdf>.
- Noack, B. (2019). Big data analytics in human resource management: Automated decision-making processes, predictive hiring algorithms, and cutting-edge workplace surveillance technologies. *Psychosociological Issues in Human Resource Management*, 7(2), 37–42.
- Wright, P. M., & Ulrich, M. D. (2017). A road well traveled: The past, present, and future journey of strategic human resource management. *Annual Review of Organizational Psychology and Organizational Behavior*, 4(1), 45–65.

DH with corpus linguistics and quantitative methods in the Humanities (e.g., in social and economic history) that are often borrowed from the social sciences.

DH thus has not only introduced computational methods to the Humanities but also significantly widened the field of inquiry, enabling new types of research, that would have been impossible to be pursued in a pre-digital age, as well as old research questions to be asked in new ways, promising to lead to new insights. Like many paradigm-shifting movements, it has sometimes been perceived in terms of culture and counter-culture and alternatively been portrayed as a “threat” or a “savior” for the Humanities as a whole.

H

Humanities (Digital Humanities)

Ulrich Tiedau
Centre for Digital Humanities, University College
London, London, UK

Big Data in the Humanities

Massive use of “Big Data” has not traditionally been a method of choice in the humanities, a field in which close reading of texts, serendipitous finds in archives, and individual hermeneutic interpretations have dominated the research culture for a long time. This “economy of scarcity” as it has been called has now been amended by an “economy of abundance,” the possibility to distance-read, interrogate, visualize, and interpret a huge number of sources that would be impossible to be read by any individual scholar in their lifetime, simultaneously by using digital tools and computational methods.

Since the mid-2000s, the latter approach is known as “Digital Humanities” (hereafter DH), in analogy to “e-Science” sometimes also as “e-Humanities,” although under the name of “Computing in the Humanities,” “Humanities Computing” or similar it has been in existence for half a century, albeit somewhat on the fringes of the Humanities canon. There are also overlaps of

Definitions of Digital Humanities

Scholars are still debating the question whether DH is primarily a set of methodologies or whether it constitutes a new discipline or is in the process of becoming a discipline, in its own right. The current academic landscape certainly allows both interpretations. Proponents of the methodological nature of DH argue that the digital turn that has embraced all branches of scholarship has not stopped for the Humanities and thus there would be no need to qualify this part of Humanities as “Digital Humanities”. In this view, digital approaches are a novel but integral part of existing Humanities disciplines, or a new version of the existing Humanities, an interpretation that occasionally even overshoots the target by equating DH with the Humanities as a whole.

On the other hand, indicators for the increasingly disciplinary character of DH are not just a range of devoted publications, e.g., *Journal of Digital Humanities* (JDH), *Digital Humanities Quarterly* (DHQ), *Literary and Linguistic Computing: the Journal of Digital Scholarship in the Humanities* (LLC), etc., and organizations, e.g., the Alliance of Digital Humanities Organizations (ADHO), an umbrella organization of five scholarly DH associations with various geographical and thematic coverage, that organizes the annual Digital Humanities conference, but also the rapid

emergence of DH centers all over the world, as this process of institutionalization following the emergence a free and novel form of inquiry is how all academic disciplines came into being originally. Melissa Terras (2012) counts 114 physical centers in 24 countries that complement long-established pioneering institutions at, e.g., the University of Virginia, the University of Maryland, and George Mason University in the USA, at the University of Victoria in Canada, or at the University of Oxford, King's College London and, somewhat newer, University College London and the University of Sussex in the UK.

History of the Field

The success of the World Wide Web and the pervasion of academia as well as everyday life by modern information and communication technology, including mobile and tablet devices, which has led to people conducting a good part of their lives online, are part of the explanation for the rapid and pervasive success of DH in academia. Against this wider technological and societal background, as well as a new iteration of the periodically recurring crises of the Humanities, a trigger for the rapid rise of the field has been its serendipitous rebranding to “Digital Humanities,” a term that has caught on widely. Whereas “Humanities Computing” emphasized the computational nature of the subject, thus tools and technology, and used “Humanities” only to qualify “Computing,” “Digital Humanities” has reversed the order, placing the emphasis firmly on the humanistic nature of the inquiry and subordinating technology to its character, thus appealing to a great number of less technologically orientated Humanities scholars. Kathleen Fitzpatrick (2011) recounts the decisive moment when Susan Schreibman, Ray Siemens, and John Unsworth, the editors of the then planned *Blackwell Companion to Humanities Computing* countered the publisher's alternative title suggestion *Companion to Digitized Humanities* with the final *Companion to Digital Humanities* (2004) because the field extended far beyond mere digitization. The name has stuck ever since, helping to bring about

a paradigmatic shift in the Humanities, which has quickly been followed by changing funding regimes, of which the establishment of an Office of Digital Humanities (2006) by the National Endowment for the Humanities (NEH) may serve as an example here.

The origins of DH can be traced all the way back to the late 1940s when the Italian priest Roberto Busa S. J., who is generally considered to be the founder of the subject and in whose honor the Alliance of Digital Humanities Organizations (ADHO) awards the annual Busa Prize, in conjunction with IBM, started working on the *Index Thomisticus*, a digital search tool for the massive corpus of Thomas Aquina's works (11 million words of medieval Latin), originally with punch card technology, resulting in a 52 volume scholarly edition that was finally published in the 1970s. The Hidden Histories project reconstructs the subject's history, or prehistory, from these times to the present with an oral history approach (Nyhan et al. 2012).

Subfields of Digital Humanities

Given its origins and the textual nature of most of the Humanities disciplines, it is no wonder that textual scholarship has traditionally been at the heart of DH, although the umbrella term also includes non-textually based digital scholarship as well. Especially in the USA, DH programs frequently developed in English departments (Kirschenbaum 2010), and the 2009 Convention of the Modern Language Association (MLA) in Philadelphia is widely seen as the breakthrough moment, at which DH became “mainstream.”

Still being a field with comparatively low maturity, there also is no clear distinction between scholars and practitioners of DH, e.g., in the heritage sector. In fact some of the most eminent work has come from, and continues to be done in, the world of libraries, archives, museums, and other heritage institutions. Generally speaking digitization has characterized a good part of early DH work, throughout the second half of the 1990s and early 2000s, creating the basis for later interpretative work, before funding bodies, in

a move to justify their investment, shifted their emphasis to funding research that utilized previously digitized data (e.g., JISC in the UK or the joint US/Canadian/British/Dutch “Digging into Data” program) rather than accumulate more digitized material that initially remained somewhat underused. Commercial companies, first and foremost Google, have been another player and contributed to the development of DH, i.e., by providing Big Humanities Data on unrivalled scales, notably Google Books with its more than 30 million digitized books (2013), although Google readily admits that this still only amounts to a fraction of all books ever published, and including integrated analytical tools like the Google n-gram viewer.

An important place in the development of digital textual scholarship has the Text Encoding Initiative (TEI). Growing from research into hypertext and the need for a standard encoding scheme for scholarly editions, this format of XML is one of the foremost achievements of early humanities computing (1987–). Apart from text encoding, digital editing, textual analytics, corpus linguistics, text mining, and language processing, central nontextual fields of activities of DH include digital image processing (still and moving), geo-spatial information systems (GIS) and mapping, data visualization, user and reader studies, social media studies, crowdsourcing, 3D/4D scanning, digital resources, subject-specific databases, web-archiving and digital long-term preservation, the semantic web, open access, and open educational practices, to name but the most important. An agreed upon “canon” of what constitutes DH next to its core of digital textual scholarship, although emerging, does not yet exist and interpretations differ in this fluid field.

Controversies and Debates

DH’s tendency to present itself, rightfully or wrongfully, as revolutionary has not only made itself friends, and like any revolutionary movement, the field also encounters criticism and backlashes. Critics, while often acknowledging the potential of DH as auxiliary means for “further knowing the

already known” or by providing shiny visualizations of research for public engagement, also question what added value “asking old research questions in new ways” has and point to the still comparatively few projects that are driven first and foremost by a humanistic research question, rather than by tool development, proof of concept, etc. Well-noted critiques came from, e.g., Stanley Fish in a series of *New York Times* columns in 2011/2012 and from Adam Kirsch in the *New Republic* in 2014, sparking ongoing controversies ever since.

Still, it is early days, and these projects are underway. As the field is maturing, it promises to transform scholarship to an even greater measure than it has already done. It does so not the least by a second novel element that DH has introduced to the Humanities with its traditionally dominant “lone-scholar ideal”: a research culture resembling the more team-based type of research done in science, technology, engineering, and medicine (STEM) subjects, in other words collaboration, an innovation just as important and potentially transformative as the use of computational methods for humanistic enquiry. DH can thus also be seen as an attempt at ending the separation of the “two cultures” in academia, the Humanities on the one hand and the STEM-subjects on the other, a notion that C. P. Snow first suggested in the 1950s. Programmatically, the final report of the first “Digging into Data” program, a collaborative funding program by US, Canadian, British, and Dutch funding bodies, bears the title “One Culture” (Wilford and Henry 2012).

Cross-References

- [Big Humanities Project](#)
- [Curriculum, Higher Education, Humanities](#)
- [Visualization](#)

Further Reading

- Berry, D. M. (Ed.). (2012). *Understanding digital humanities*. Basingstoke: Palgrave MacMillan.
- Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital humanities*. Cambridge, MA: MIT Press.

- Fish, S. (2011, December 26). The old order changeth. *Opinionator Blog, New York Times*. <http://opinionator.blogs.nytimes.com/2011/12/26/the-old-order-changeth/>. Accessed August 2014.
- Fitzpatrick, K. (2011). The humanities done digitally. *The Chronicle of Higher Education*. <http://chronicle.com/article/The-Humanities-Done-Digitally/127382/>. Accessed August 2014.
- Gold, M. (Ed.). (2012). *Debates in the digital humanities*. Minneapolis: Minnesota University Press.
- Kirsch, A. (2014, May 2). Technology is taking over English departments: The false promise of the digital humanities. *New Republic*. <http://www.newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch>. Accessed August 2014.
- Kirschenbaum, M. G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin*, 150, 1–7.
- McCarty, W. (2005). *Humanities computing*. Basingstoke: Palgrave.
- Nyhan, J., Flynn, A., & Welsh, A. (2012). A short introduction to the Hidden Histories project. *Digital Humanities Quarterly*, 6(3). <http://www.digitalhumanities.org/dhq/vol/6/3/000130/000130.html>. Accessed August 2014.
- Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). *A companion to digital humanities*. Oxford: Blackwell.
- Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2007). *A companion to digital literary studies*. Oxford: Blackwell.
- Terras, M. (2012). *Infographic: Quantifying digital humanities*. <http://blogs.ucl.ac.uk/dh/2012/01/20/infographic-quantifying-digital-humanities/>. Accessed August 2014.
- Terras, M., Nyhan, J., & Vanhoutte, E. (Eds.). (2013). *Defining digital humanities: A reader*. Farnham: Ashgate. ISBN 978-1-4094-6963-6.
- Warwick, C., Terras, M., & Nyhan, J. (Eds.). (2012). *Digital humanities in practice*. London: Facet.
- Wiliford, C., & Henry, C. (2012). *One culture: Computationally intensive research in the humanities and social sciences. A report on the experiences of first respondents to the digging into data challenge* (CLIR Publication No. 151). Washington, DC: Council on Library and Information Resources.

Indexed Web, Indexable Web

► [Surface Web vs Deep Web vs Dark Web](#)

Indicator Panel

► [Dashboard](#)

Industrial and Commercial Bank of China

Jing Wang¹ and Aram Sinnreich²

¹School of Communication and Information,
Rutgers University, New Brunswick, NJ, USA

²School of Communication, American University,
Washington, DC, USA

The Industrial and Commercial Bank of China (ICBC)

The Industrial and Commercial Bank of China (ICBC) was the first state-owned commercial bank of the People's Republic of China (PRC). It was founded on January 1st, 1984, and is headquartered in Beijing. In line with Deng Xiaoping's economic reform policies launched in the late 1970s, the State Council (chief administrative authority of China) decided to relay all

the financial businesses related to industrial and commercial sectors from the central bank (People's Bank of China) to ICBC (China Industrial Map Committee 2016). This decision made in September 1983 is considered a landmark event in the evolution of China's increasingly specialized banking system (Fu and Hefferman 2009). While the government retains control over ICBC, the bank began to take on public shareholders in October, 2006. As of May 2016, ICBC was ranked as the world's largest public company by *Forbes* "Global 2000." (Forbes Ranking 2016) With its combination of state and private ownership, state governance, and commercial dealings, ICBC serves as a perfect case study to examine the transformation of China's financial industry.

Big data collection and database construction are fundamental to ICBC's management strategies. Beginning in the late 1990s, ICBC paid unprecedented attention on the implication of information technology (IT) in their daily operations. Several branches adopted computerized input and internet communication of transactions, which had previously relied upon manual practices by bank tellers. Technological upgrades increased work efficiency and also helped to save labor costs. More importantly, compared to the labor-driven mechanism, the computerized system was more effective for retrieving data from historical records and analyzing these data for business development. At the same time, it became easier for the headquarters to control the local branches by checking digitalized

information records. Realizing the benefits of these informatization and centralization tactics, the head company assigned its Department of Information Management to develop a centralized database collecting data from every single branch. This database is controlled and processed by ICBC headquarters but is also available for use by local branches with the permission of top executives.

In this context, “big data” refers to all the information collected from ICBC’s daily operations and can be divided into two general categories: “structured data” (which is organized according to preexisting database categories) and “unstructured data” (which does not) (Davenport and Kim 2013). For example, a customer’s account information is typically structured data. The branch has to input the customer’s gender, age, occupation, etc., into the centralized network. This information then flows into the central database which is designed specifically to accommodate it. Any data other than the structured data will be stored as raw data and preserved without processing. For example, the video recorded at a local branch’s business hall will be saved with only a date and a location label. Though “big data” in ICBC’s informational projects refers to both structured and unstructured data, the former is the core of ICBC’s big data strategy and is primarily used for data mining.

Since the late 1990s, ICBC has invested in big data development with increasingly large economic and human resources. On September 1st, 1999, ICBC inaugurated its “9991” project, which aimed at centralizing the data collected from ICBC branches nationwide. This project took more than 3 years to accomplish its goal. Beginning in 2002, all local branches were connected to ICBC’s Data Processing Center in Shanghai – a data warehouse with a 400 terabyte (TB) capacity. The center’s prestructured database enables ICBC headquarters to process and analyze data as soon as they are generated, regardless of the location. With its enhanced capability in storing and managing data, ICBC also networked and digitized its local branch operations. Tellers are able to input

customer information (including their profiles and transaction records) into the national Data Center through their computers at local branches. These two-step strategies of centralization and digitization allow ICBC to converge local operations on one digital platform, which intensifies the headquarters’ control over national businesses. In 2001, ICBC launched another data center in Shenzhen, China, which is in charge of the big data collected from its overseas branches. ICBC’s database thus enables the headquarters’ control over business and daily operations globally and domestically.

By 2014, ICBC’s Data Center in Shanghai had collected more than 430 million individual customers’ profiles and more than 600,000 commercial business records. National transactions – exceeding 215 million on daily basis – have all been documented at the Data Center. Data storage and processing on such a massive scale cannot be fulfilled without a powerful and reliable computer system. The technology infrastructure supporting ICBC’s big data strategy consists of three major elements: hardware, software, and cloud computing. Suppliers are both international and domestic, including IBM, Teradata, and Huawei.

Further, ICBC has also invested in data backup to secure its database infrastructure and data records. The Shanghai Data Center has a backup system in Beijing which can record data when the main server fails to work properly. The Beijing data center serves as a redundant system in case the Shanghai Data Center fails. It only takes less than 30 s to switch between two centers. To speed data backup and minimize data loss in significant disruptive events, ICBC undertakes multiple disaster recovery (DR) tests on a regular basis.

The accumulation and construction of big data is significant for ICBC’s daily operation in three respects. First of all, big data allows ICBC to develop its customers’ business potential through a so-called “single-view” approach. A customer’s business data collected from one of ICBC’s 35 departments are available for all the other departments. By mining the shared database, ICBC headquarters is able to evaluate both a

customer's comprehensive value and the overall quality of all existing customers. Cross departmental business has also been propelled (e.g., the Credit Card Department may share business opportunities with the Savings Department). Second, the ICBC marketing department has been using big data for email-based marketing (EBM). Based on the data collected from branches, the Marketing and Business Development Department is able to locate their target customers and follow up with customized marketing/advertising information via customized email communications. This data-driven marketing approach is increasingly popular among financial institutions in China. Third, customer management systems rely directly on big data. All customers have been segmented into six levels, ranging from "one star" to "seven stars," (one star and two stars fall into a single segment which indicates the customers' savings or investment levels at ICBC). "Seven Stars" clients have the highest level of credit and enjoy the best benefits provided by ICBC.

Big data has influenced ICBC's decision-making on multiple levels. For local branches, market insights are available at a lower cost. Consumer data generated and collected at local branches have been stored on a single platform provided and managed by the national data center. For example, a branch in an economically developing area may predict demand for financial products by checking the purchase data from branches in more developed areas. The branch could also develop greater insights regarding the local consumer market by examining data from multiple branches in the geographic area. For ICBC headquarters, big data fuels a dashboard through which it monitors ICBC's overall business and is alerted to potential risks. Previously, individual departments used to manage their financial risk through their own balance sheets. This approach was potentially misleading and even dangerous for ICBC's overall risk profile. A given branch providing many loans and mortgages may be considered to be performing well, but if a large number of branches overextended themselves, the

emergent financial consequences might create a crisis for ICBC or even for the financial industry at large. Consequently, today, a decade after its data warehouse construction, ICBC considers big data indispensable in providing a holistic perspective, mitigating risk for its business and development strategies.

To date, ICBC has been a pioneer in big data construction among all the financial enterprises in China. It was the first bank to have all local data centralized in a single database. As the Director of ICBC's Informational Management Department claimed in 2014, ICBC has the largest Enterprise Database (EDB) in China.

Parallel to its aggressive strategies in big data construction, the issue of privacy protection has always been a challenge in ICBC's customer data collection and data mining. The governing policies primarily regulate the release of data from ICBC to other institutions, yet the protection of customer privacy within ICBC itself has rarely been addressed. According to the central bank's *Regulation on the Administration of the Credit Investigation Industry* issued by the State Council in 2013, interbank sharing of customer information is forbidden. Further, a bank is not eligible to release customer information to its nonbanking subsidiaries. For example, the fund management company (ICBCCS) owned by ICBC is not allowed access customer data collected from ICBC banks. The only situation in which ICBC could release customer data to a third party is when such information has been linked to the official inquiry by law enforcement. These policies prevent consumer information from leaking to other companies for business purposes. Yet, the policies have also affirmed the fact that ICBC has full ownership of the customer information, thus giving ICBC greater power to use the data in its own interests.

Cross-References

- [Data Mining](#)
- [Data Warehouse](#)
- [Structured Data](#)

Further Reading

- China Industrial Map Editorial Committee, China Economic Monitoring & Analysis Center & Xinhua Holdings. 2016. *Industrial map of China's financial sectors*, Chapter 6. World Scientific Publishing.
- Davenport, T., & Kim, J. (2013). *Keeping up with the quants: Your guide to understanding and using analytics*. Boston: Harvard Business School Publishing.
- Fu, M., & Hefferman, S. (2009). The effects of reform on China's bank structure and performance. *Journal of Banking & Finance*, 33(1), 39–52.
- Forbes Ranking (2016). The World's Biggest Public Company. Retrieved from <https://www.forbes.com/companies/icbc/>.

Informatics

Anirudh Prabhu
Tetherless World Constellation, Rensselaer
Polytechnic Institute, Troy, NY, USA

Synonyms

Information Engineering; Information Science; Information Systems; Information Studies; Information Theory; Informatique

Definition

Informatics is the science of information, the practice of information processing, and the engineering of information systems. Informatics studies the structure, algorithms, behavior, and interactions of natural and artificial systems that store, process, access, and communicate information (Xinformatics Concept 2012).

The advent of “big data” has brought many opportunities for people and organizations to leverage large amounts of data to answer previously unanswered questions, but along with these opportunities come problems of storing and processing these data. Expertise in the field of Informatics becomes essential to build new information systems or adapt existing information systems to address “big data.”

History

The French term “Informatique” was coined in March 1962 by Phillipe Dreyfus – along with translations in various other languages. Simultaneously and independently Walter Bauer and his associates proposed the English term “Informatics” when they co-founded Informatics Inc. (Fourman 2002). A very early definition of “Informatics” from Mikhailov in 1967 states that “Informatics is the discipline of science which investigates the structure and properties (not specific content) of scientific information, as well as the regularities of scientific information activity, its theory, history, methodology and organization” (Fourman 2002). But in recent times, the scope of Informatics has moved way beyond just scientific information. It now extends to all information in the modern age.

Need for “X-Informatics”

The word Informatics is used as a compound, in conjunction with the name of a discipline, for example, Bioinformatics, Geoinformatics, Astroinformatics. Earlier, people who had deep knowledge of a specific domain would work on processing and engineering information systems that would be designed only for that domain.

In the last decade, fueled by the rapid increase in data and information resources, Informatics has gained greater visibility across a broad range of disciplines. As the popularity of Informatics increased through time, there has been a widespread need for people who specialized in the field of X-informatics. Informaticians (or Informaticists) and Data Scientists are trained in Informatics Theory, which is a combination of information science, cognitive science, social science, library science, and computer science, enables these people to engineer information systems in various domains using the Informatics methodologies. In the term X-informatics, “X” is a variable for the domain, which can be bio, geo, chem, astro, etc. The term indicates that knowledge of Informatics can be applied to many

different domains. Hence, there are many academic institutions across the world which have specialized courses and even degrees in Informatics.

Informatics in Data-Information-Knowledge Ecosystem

The amount of data in the world is exponentially rising. But these data are not directly useful to the majority of people. In order for the data to be used to its fullest potential, it needs to be processed, represented, and communicated in a meaningful way. This is where Informatics methods come into the picture. Figure 1 shows the Data-Information-Knowledge ecosystem and the focus of Informatics methods in this ecosystem. Informatics methods focus on transforming raw data into information that can easily understood. Once meaningful information has reached the “consumer,” they can draw inferences, have conversations, combine the new information with previous experiences, and gain knowledge on a specific subject.

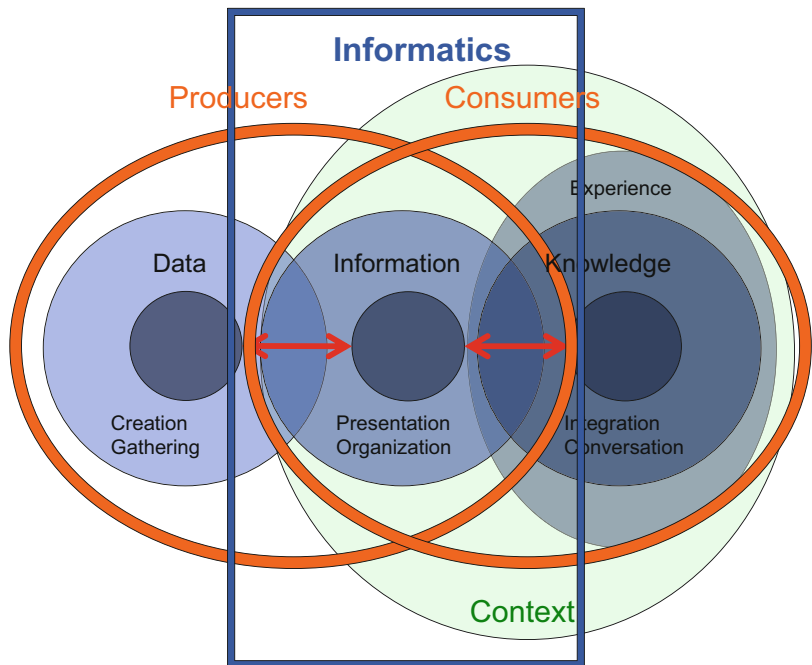
Concepts in Informatics

Data – Data are encodings that represent the qualitative or quantitative attributes of a variable or a set of variables. Data are often viewed as the lowest level of abstraction form which information and knowledge are derived (Fox 2016).

Information – Representations of “Data” in a form that lends itself to human use. Information has three indivisible ingredients – content, context, and structure (Fox 2016).

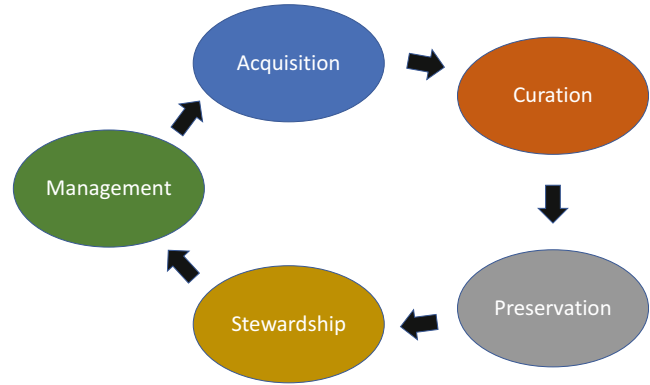
Information Theory – “Information theory is the branch of mathematics that describes how uncertainty should be quantified, manipulated and represented” (Ghahramani 2003). Information theory is one of the rare scientific fields to have an identifiable origin. It was originally proposed by Claude E. Shannon in 1948 in a landmark paper titled “A Mathematical Theory of Communication.” In this paper, “information” can be considered as a set of messages, where the goal is to send this “information” over a noisy channel, and then to have the receiver reconstruct the message with a low probability of error.

Informatics, Fig. 1 Data-Information-Knowledge ecosystem (Fox 2016)



Informatics,

Fig. 2 Information life cycle



Information Entropy – Information entropy is defined as the average amount of information produced by a probabilistic stochastic source of data. Mathematically, Information Entropy can be defined as,

$$H = - \sum_{i=1}^n p_i \times \log_2(p_i)$$

where, H is the “Entropy” and p_i is the probability of occurrence of the i -th possible value of the source message. Information Entropy is commonly measured in “bits,” which represents 2 possible states. Hence, base of the logarithm in the definition of entropy is “2.” If the entropy measurement unit changes, then the base of the logarithm also changes accordingly. For example, if information entropy is measure in decimal digits, then the base of the logarithm would change to “10.”

“In information theory, entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process” (Ghahramani 2003). Therefore, Entropy is a key measure in informatics.

Information Architecture – Information Architecture is the art of expressing a model or concept of information used in activities that require explicit details of complex systems. Richard Saul Wurman an architect and a graphic design, popularized the usage of the term. Wurman defines an information architect as follows – “. . . I mean architect as in the creating of systemic, structural, and orderly principles to

make something work – the thoughtful making of either artifact, or idea, or policy that informs because it is clear” (Fox 2016).

Information Life Cycle – Information Life Cycle refers to the steps stored information goes through from its creation to its deletion or archival (Fig. 2).

Stages of the Information Life Cycle are as follows (Fox 2016):

Acquisition: The process of recording or generating a concrete artifact from the concept

Curation: The activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future

Preservation: The process of retaining usability of data in some source form for intended and unintended use

Stewardship: The process of maintaining integrity across acquisition, curation, and preservation

Management: The process of arranging for discovery, access and use of data, information and all related elements. Management also includes overseeing processes for acquisition, curation, preservation, and stewardship

Informatics in Scientific Research

As mentioned earlier, in the past, Informatics efforts emerged largely in isolation across a number of disciplines. “Recently, certain core elements in informatics have been recognized as being applicable across disciplines. Prominent domains of informatics have two key factors in

common: i) a distinct shift toward methodologies and away from dependence on technologies and ii) understanding the importance of and thereby using multidisciplinary and collaborative approaches in research” (Fox 2011).

As a domain, Informatics builds on several existing academic disciplines, primarily Artificial Intelligence, Cognitive Science and Computer Science. Cognitive Science concerns the study of natural information processing systems; Computer Science concerns the analysis of computation, and the design of computing systems; Artificial Intelligence plays a connecting role, producing systems designed to emulate those found in nature (Fourman 2002).

Example Use case

The “Co-Evolution of the Geo- and Biosphere” project will be used as the example use case for exhibiting Informatics techniques. The main goal of this project is to use known informatics techniques in diverse disciplines (like mineralogy, petrology, paleo biology, paleo tectonics, geochemistry, proteomics, etc.) to discover patterns in evolution of Earth’s environment that exemplifies the abovementioned “Co-evolution.”

There are vast amounts of data available in each scientific discipline. Not all of them are

immediately useable for analysis. The first step is to process the information into a format than can be used for modeling and visualizing the data.

For example, the most commonly used databases in mineralogy are “mindat.org” and “RRUFF.info.” Combined, these databases contain data on all the mineral localities on Earth. Along with data on the localities, they also contain data on the different minerals, their chemical composition, age of minerals, and other geologic properties. Since one of the goals is to discover pattern and trends in the evolution of Earth’s environment, minerals that most often occurred together need to be observed and analyzed. One of the best ways to **represent** this co-occurrence information is to create a network visualization where every node represents a mineral and the edges (or connections) implies that two minerals occur at the same locality. To create this visualization, the raw data need to be **processed** into the required format. Adjacency matrices and edge lists are appropriate formats for a network structure. In Tables 1 and 2, the difference between the “Raw data” and the “Processed Data” can be seen.

Once the data have been converted to the required format, the network visualization can be created. It is important that the visualization created **communicates** the intended information (in this case the Manganese Mineral Environment) to

Informatics, Table 1 Raw data – manganese minerals from the website “RRUFF.info”

Names	RRUFF ID	Ideal chemistry	Source	Locality
Akatoreite	R060230	$\text{Mn}^{2+}_9\text{Al}_2\text{Si}_8\text{O}_{24}(\text{OH})_8$	Michael Scott S100146	Near mouth of Akatore Creek, Taieri, Otago Province, New Zealand
Akrochordite	R100028	$\text{Mn}^{2+}_5(\text{AsO}_4)_2(\text{OH})_4 \cdot 4\text{H}_2\text{O}$	William W. Pinch	Langban, Filipstad, Varmland, Sweden
Alabandite	R070174	MnS	Michael Scott S101601	Mina Preciosa, Sangue de Cristo, Puebla, Mexico
Allactite	R070175	$\text{Mn}^{2+}_7(\text{AsO}_4)_2(\text{OH})_8$	Michael Scott S102971	Langban, Filipstad, Varmland, Sweden
Allactite	R150120	$\text{Mn}^{2+}_7(\text{AsO}_4)_2(\text{OH})_8$	Steven Kuitems	Sterling Mine, 1200' Level, Ogdensburg, New Jersey, USA
Alleghanyite	R060904	$\text{Mn}^{2+}_5(\text{SiO}_4)_2(\text{OH})_2$	Michael Scott S100995	Near Bald Knob, Alleghany County, North Carolina, USA

Informatics, Table 2 Processed data – co-occurrence edge list for manganese minerals

Source	Target	Value
Agmantinite	Alabandite	0.1
Akatoreite	Alabandite	0.75
Akhtenskite	Alabandite	0.5
Akrochordite	Alabandite	0.8
Akrochordite	Allactite	0.5
Alabandite	Allactite	0.846153846
Akhtenskite	Allegghanyite	0.333333333
Akrochordite	Allegghanyite	0.6
Alabandite	Allegghanyite	0.632183908
Allactite	Allegghanyite	0.461538462
Alabandite	Alluaivite	0.75
Alabandite	Andrianovite	0.666666667
Alluaivite	Andrianovite	0.666666667
Alabandite	Ansermetite	0.666666667
Allegghanyite	Ansermetite	0.888888889

the “audience.” For example, a force directed network not only indicates which nodes are connected to each other, it also indicates the most connected nodes (node size), the stable geometric layout with similar nodes in the network exhibiting closer proximity to each other, all the nodes can be grouped on many different properties, thereby also indicating how each group behaves in the network environment. Interactive 2D and 3D Manganese Networks (with 540 mineral nodes) can be found at (https://dtdi.carnegiescience.edu/sites/all/themes/bootstrap-d7-theme/networks/Mn/network/Mn_network.html) and (https://deeptime.tw.rpi.edu/viz/3D_Network/Mn_Network/index.html). These visualizations show a force directed layout with nodes groups by Oxidation state (Indicator of loss of electrons of an atom in a chemical compound), Paragenetic Modes (Formational conditions of the mineral), or Mineral Age. With these visualizations, the “audience” can explore some complex and hidden patterns of diversity and distribution in mineral environment.

Conclusion

As digital data continue to increase at unprecedented rate, importance of informatics

applications increases. It is for this very reason that many universities across the world (especially in the USA) have started multidisciplinary informatics programs. These programs give students the freedom to apply their knowledge of informatics to their field of choice. With “big data” becoming increasingly common in most disciplines, informatics as a domain will not be losing traction anytime soon.

Further Reading

Fourman, M. (2002). Informatics. In *International encyclopedia of information and library science*. London, UK: Routledge.

Fox, P. (2011). The rise of informatics as a research domain. In *Proceedings of WIRADA Science Symposium*, Melbourne (Vol. 15, pp. 125–131).

Fox, P. (2016). Tetherless world constellation. Retrieved 22 Sept 2017, from <https://tw.rpi.edu/web/courses/xinformatics/2016>.

Ghahramani, Z. (2003). Information theory. In *Encyclopedia of cognitive science*. London, UK: Nature Publishing Group.

Xinformatics Concept. (2012). Tetherless World Constellation. Retrieved 22 Sept 2017, from <https://tw.rpi.edu/web/concept/XinformaticsConcept>.

**Information Commissioner,
United Kingdom**

Ece Inan
Girne American University Canterbury,
Canterbury, UK

The Information Commissioner’s Office (ICO) is the UK’s independent public authority which is responsible for data protection mainly in England, Scotland, Wales, and Northern Ireland; and also ICO has right to conduct some international duties. ICO was firstly set up to uphold information rights by implementing the Data Protection Act 1984. The ICO declared their mission statement as to promote respect for the private lives of individuals and in particularly, for the privacy of their information by implementing the Data Protection Act 1984 and also influencing national and

international thinking on privacy and personal information.

ICO enforces and oversees all the data protection issues by following the Freedom of Information Act 2000, Environmental Information Regulations 2004, and Privacy and Electronic Communications Regulations 2003, and also ICO has some limited responsibilities under the INSPIRE Regulations 2009, in England, Wales, Northern Ireland, and UK-wide public authorities based in Scotland. On the other hand, Scotland has complementary INSPIRE Regulations and its own Scottish Environmental Information Regulations regulated by the Scottish Information Commissioner and the Freedom of Information (Scotland) Act 2002.

The Information Commissioner is appointed by the Queen and reports directly to Parliament. The Commissioner is supported by the management board. The ICO's headquarter is in Wilmslow, Cheshire; in addition to this, three regional offices in Northern Ireland, Scotland, and Wales are aimed to provide relevant services where legislation or administrative structure is different.

Under the Freedom of Information Act, Environmental Information Regulations, INSPIRE Regulations, and associated codes of practice, the functions of the ICOs contain noncriminal enforcement and assessments of good practice, providing information to individuals and organizations, taking appropriate action when the law or freedom of information is broken, considering complaints, disseminating publicity and encouraging sectoral codes of practice, and taking action to change the behavior of organizations and individuals that collect, use, and keep personal information. The main aim is to promote data privacy for individuals, for providing this service, the ICO has different tools such as criminal prosecution, noncriminal enforcement, and audit. The Information Commissioner also has the power to serve a monetary penalty notice on a data controller and promotes openness to public.

The Data Protection Act 1984 introduced basic rules of registration for users of data and rights of access to that data for the individuals to which it

related. In order to comply with the Act, a data controller must comply with the following eight principles as "data should be processed fairly and lawfully; should be obtained only for specified and lawful purposes; should be adequate, relevant, and not excessive; should be accurate and, where necessary, kept up to date; should not be kept longer than is necessary for the purposes for which it is processed; should be processed in accordance with the rights of the data subject under the Act; should be appropriate technical and organisational measures should be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data; and should not be transferred to a country or territory outside the European Economic Area unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to the processing of personal data."

In 1995, The EU formally adopted the General Directive on Data Protection. In 1997, DUIS, the Data User Information System, was implemented, and the Register of Data Users was published on the internet. In 2000, the majority of the Data Protection Act comes into force. The name of the office was changed from the Data Protection Registrar to the Data Protection Commissioner. Notification replaced the registration scheme established by the 1984 Act. Revised regulations implementing the provisions of the Data Protection Telecommunications Directive 97/66/EC came into effect. In January 2001, the office was given the added responsibility of the Freedom of Information Act and changed its name to the Information Commissioner's Office. On 1 January, 2005, the Freedom of Information Act 2000 was fully implemented. The Act was intended to improve the public's understanding of how public authorities carry out their duties, why they make the decisions they do, and how they spend their money. Placing more information in the public domain would ensure greater transparency and trust and widen participation in policy debate. In October 2009, the ICO adopted a new mission statement: "The ICO's mission is to uphold

information rights in the public interest, promoting openness by public bodies and data privacy for individuals.” In 2011, ICO launched the “data sharing code of practice” at the House of Commons and enable to impose monetary penalties of up to £500,000 for serious breaches of the Privacy and Electronic Communications Regulations.

Cross-References

- [Open Data](#)

Further Reading

Data Protection Act 1984. <http://www.out-law.com/page-413>. Accessed Aug 2014.

DataProtectionAct 1984. http://www.legislation.gov.uk/ukpga/1984/35/pdfs/ukpga_19840035_en.pdf?view=extent. Accessed Aug 2014.

Smartt, U. (2014). *Media & entertainment law* (2nd ed.). London: Routledge.

Information Discovery

- [Data Discovery](#)
- [Data Processing](#)

Information Engineering

- [Informatics](#)

Information Extraction

- [Data Processing](#)

Information Hierarchy

- [Data-Information-Knowledge-Action Model](#)

Information Overload

Deepak Saxena¹ and Sandul Yasobant²

¹Indian Institute of Public Health Gandhinagar, Gujarat, India

²Center for Development Research (ZEF), University of Bonn, Bonn, Germany

Background

With the advent of technology, humans are now afforded greater access to information than ever before (Lubowitz and Poehling 2010) and many can have access to any information irrespective of its relevance. However, evidence indicates that humans have a limited capacity to process and retain new information (Lee et al. 2017; Mayer and Moreno 2003). This capacity is influenced by multiple personal factors such as anxiety (Chae et al. 2016), motivation to learn, and existing knowledge base (Kalyuga et al. 2003). Information overload occurs when the volume or complexity of information accessed by an individual exceeds their capacity to process the information within a given timeframe (Eppler and Mengis 2004; Miller 1956).

History of Information Overload

The term “information overload” has been existence for more than 2000 years. This has been re-emerging as a new phenomenon in the recent digital world. Since the introduction of the printing machine in Europe in the fifteenth century to the current millions of Google search on the Internet, the problem of information overload remained a conundrum (Blair 2011).

Definition

Although a user-friendly definition of information overload is still missing; Roetzel (2018), contributed a working definition.

Information overload is a state in which a decision-maker faces a set of information (i.e., an information load with informational characteristics such as

an amount, a complexity, and a level of redundancy, contradiction and inconsistency) comprising the accumulation of individual informational cues of differing size and complexity that inhibit the decision maker's ability to optimally determine the best possible decision. The probability of achieving the best possible decision is defined as decision-making performance. The suboptimal use of information is caused by the limitation of scarce individual resources. A scarce resource can be limited individual characteristics (such as serial processing ability, limited short-term memory) or limited task-related equipment (e.g., time to make a decision, budget).

Information Overload: Double-Edged Sword: Problem or Opportunity?

The simplicity of creating, duplicating, and sharing information online in high volumes resulted in the information overload. The most cited causes of information overload are the existence of multiple sources of information, over-abundance of information, difficulty in managing information, irrelevance/unimportance of the received information, and scarcity of time on the part of information users to analyze and understand information (Eppler and Mengis 2004).

The challenge is how to alleviate the burden of information. As there is no thumb rule for this, keeping things simple, relevant, clear, straight forward makes a step towards the reduction of overload. As per Blair, who identified four "S's for managing information overload is: *storing, sorting, selecting, and summarizing*" (Morrison 2018). One mystery raised by the issue of information overload is that infinitely increasing both information and the capacity to use that information does not guarantee better decisions leading to desired outcomes, which is somehow not true. After all, information is often irrelevant, because either people are simply set in their ways, or natural and social systems are too unpredictable, or people's ability to act is somehow restrained. What is required, then, is not just a skill in prioritizing information, but an understanding of when information is not needed. In real phenomenon, information overload might prevent taking the right decision or action because of its nature of huge volume;

however, with careful use, it could be managed for the right policy decision.

Specialists agree that, for information users and information professionals alike, achieving information literacy is vital for successfully dealing with information overload (Bruce 2013). Information literacy as per Edmunds et al. has been defined as "a set of abilities requiring individuals to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information" (Edmunds and Morris 2000). An information literate person can determine the extent of information, access the need for information, evaluate it, incorporate and use it effectively (Gausul Hoq 2014). The scholarships indicate that, to judiciously use the information from various sources for problem-solving, a person should acquire at least a moderate level of information literacy. Admittedly, this is not an easy task and even the most expert information seekers could be overwhelmed by the huge quantity of information from which to find his/her required information. However, as one continues acquiring, upgrading and refining information literacy skills, he/she will find it easier to deal with information overload in the long run (Benselin and Ragsdell 2016).

Conclusion

The overload of information that has been experienced today as millions of Google search results in a fraction of a second surely can be a privilege, which results into the massively increased access to the consumption and production of information in the digital age but which one to utilize, absorb, and imbibe is difficult. Although the information overload creates problems, it has also inspired important solutions for evidence generation. The foregoing discussions have made it clear that the problem of information overload is here to stay and with a growing focus on research and development in the coming decade, its intensity will only increase. With the advent of new technologies and various techniques of self-publishing, information overload will surely present itself to a worldwide audience in new shapes and

dimensions in the near future. There might be great potential for the policymakers to use this information overload in a positive way in the process of evidence-based policy formulation. Although the quality of life is greatly influenced by the information overload in either way, the ease of accessing information with in fraction of second need to be considered as the positive aspect of it. However, it depends on the user, who is accessing this huge information, and the decision-capacity and the knowledge level of the user to use this effectively and efficiently.

Further Reading

- Benselin, J. C., & Ragsdell, G. (2016). Information overload: The differences that age makes. *Journal of Librarianship and Information Science*, 48(3), 284–297. <https://doi.org/10.1177/0961000614566341>.
- Blair, A. (2011). Information overload's 2,300-year-old history. *Harvard Business Review*, 1. Retrieved from <https://hbr.org/2011/03/information-overloads-2300-yea.html>.
- Bruce, C. S. (2013). *Information literacy research and practice: an experiential perspective* (pp. 11–30). https://doi.org/10.1007/978-3-319-03919-0_2.
- Chae, J., Lee, C., & Jensen, J. D. (2016). Correlates of cancer information overload: Focusing on individual ability and motivation. *Health Communication*, 31(5), 626–634. <https://doi.org/10.1080/10410236.2014.986026>.
- Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, 20(1), 17–28. [https://doi.org/10.1016/S0268-4012\(99\)00051-1](https://doi.org/10.1016/S0268-4012(99)00051-1).
- Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society*, 20(5), 325–344. <https://doi.org/10.1080/01972240490507974>.
- Gausul Hoq, K. M. (2014). Information overload: causes, consequences and remedies: A study. *Philosophy and Progress*, LV–LVI, 49–68. <https://doi.org/10.3329/pp.v55i1-2.26390>.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4.
- Lee, K., Roehrer, E., & Cummings, E. (2017). Information overload in consumers of health-related information. *JBIS Database of Systematic Reviews and Implementation Reports*, 15(10), 2457–2463. <https://doi.org/10.1112/JBISIR-2016-003287>.
- Levitin, D. J. (2014). *The organized mind: Thinking straight in the age of information overload*. New York: Dutton. ISBN-13: 978-0525954187.
- Lubowitz, J. H., & Poehling, G. G. (2010). Information overload: Technology, the internet, and arthroscopy. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 26(9), 1141–1143. <https://doi.org/10.1016/j.arthro.2010.07.003>.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52. Retrieved from <http://faculty.washington.edu/farkas/WDFR/MayerMoreno9WaysToReduceCognitiveLoad.pdf>.
- Miller, G. A. (1956). The magical number seven, plus or minus two some limits on our capacity for processing information. *Psychological Review*, 101. Retrieved from <http://spider.apa.org/ftdocs/rev/1994/april/rev1012343.html>.
- Morrison, R. (2018). *Empires of knowledge: Scientific networks in the early modern world*. (P. Findlen, Ed.). New York: Routledge. 2019: Routledge. <https://doi.org/10.4324/9780429461842>.
- Orman, L. V. (2016). *Information overload paradox: Drowning in information, starving for knowledge*. North Charleston: CreateSpace Independent Publishing Platform. ISBN-13: 978-1522932666.
- Pijpers, G. (2012). *Information overload: A system for better managing everyday data*. Hoboken: Wiley Online Library. ISBN 9780470625743.
- Roetzel, P. G. (2018). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, 1–44. <https://doi.org/10.1007/s40685-018-0069-z>.
- Schultz, T. (2011). The role of the critical review article in alleviating information overload. *Annual Reviews*, 56. Available from: https://www.annualreviews.org/pb-assets/ar-site/Migrated/Annual_Reviews_WhitePaper_Web_2011-1293402000000.pdf.

Information Quantity

Martin Hilbert

Department of Communication, University of California, Davis, Davis, CA, USA

The question of “how much information” there is in the world goes at least back to the times when Aristotle’s student Demetrius (367 BC – ca.283 BC) was asked to organize the Library of Alexandria in order to collect and quantify “how many thousand books are there” (Aristeas 200AD, sec. 9). Pressed by the exploding number of information and communication technologies (ICT)

during recent decades, several research projects have taken up this question again since the 1960s. They differ considerably in focus, scope, and measurement variable. Some used US\$ as a proxy for information (Machlup 1962; Porat 1977), others number of words (Ito 1981; Pool 1983; Pool et al. 1984), some focused on the national level of a single country (Dienes 1986, 2010), others broad estimations for the entire world (Gantz and Reinsel 2012; Lesk 1997; Turner et al. 2014), some focused on unique information (Bounie 2003; Lyman et al. 2000), and others on a specific sector of society (Bohn and Short 2009; Short et al. 2011) (for a methodological comparison and overview see Hilbert 2012, 2015a).

The big data revolution has provided much new interest in the idea of quantifying the amount of information in the world. The idea is that an important early step in understanding a phenomenon consists in quantifying it: “when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind” (Lord Kelvin, quoted from Bartlett 1968, p. 723). Understanding the data revolution implies understanding how it grows and evolves.

In this inventory, we mainly follow the methodology of what has become a standard reference in estimating the world’s technological information capacity: Hilbert and López (2011). The total technological storage capacity is calculated as the sum of the product of technological devices and their informational storage performance. Technological performance is measured in the installed binary hardware digits, which is then normalized on compression rates. The hardware performance is estimated as “installed capacity” (not the effectively used capacity), which implies that it is assumed that all technological capacities are used to their maximum. For storage this evaluates the maximum available storage space (“as if all storage were filled”). The normalization on software compression rates is important for the creation of meaningful time series, as compression algorithms have enabled to store more information

on the same hardware infrastructure over recent decades (Hilbert 2014a; Hilbert and López 2012a). We normalize on “optimally compressed bits” (as if all content were compressed with the best compression algorithms possible in 2014 (Hilbert and López 2012b)). It would also be possible to normalize on a different standard, but the optimal level of compression has a deeper information theoretic conceptualization as it approaches the entropy of the source (Shannon 1948). For the estimation of compression rates of different content, justifiable estimates are elaborated for 7-years intervals (1986, 1993, 2000, 2007, 2014). For more see Hilbert (2015b) and López and Hilbert (2012).

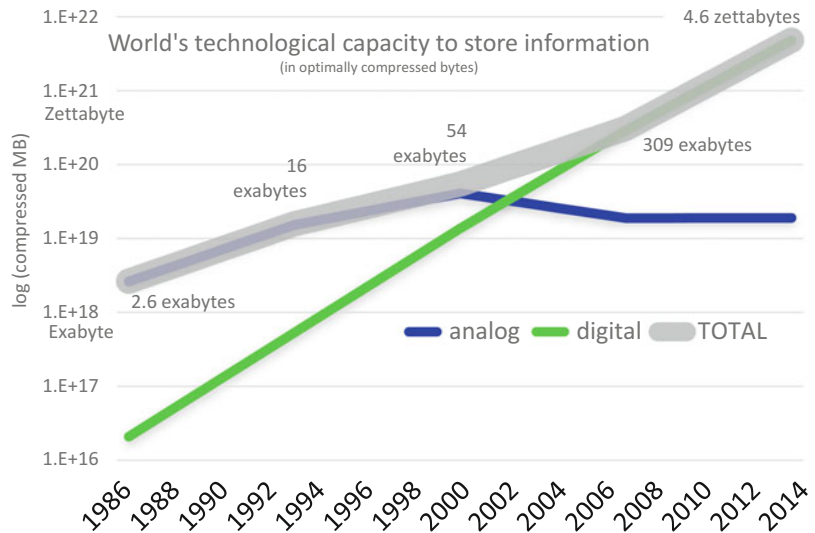
For the following result, the estimations for the period 1986–2007 follow Hilbert and López (2011). The update for 2007–2014 follows a mix of estimates, including comparisons with more current updates (Gantz and Reinsel 2012; Turner et al. 2014).

Figure 1 shows that the world’s technological capacity to store information has almost reached 5 zettabytes in 2014 (from 2.6 exabytes in 1986 to 4.6 zettabytes in 2014). This results in a compound annual growth rate of some 30%. This is about five times faster than the world economy grew during the same period. The digitalization of the world’s information stockpile happened in what is a historic blink of an eye: in 1986, less than 1% of the world’s mediated information was still stored in digital format. By 2014, less than 0.5% is stored in analog media. Some analog storage media are still growing strongly today. For example, it is well known that the long-promised “paperless office” has still not arrived. The usage of paper still grows with some 15% per year (some 2.5 times faster than the economy), but digital storage is growing at twice that speed. The nature of this exponential growth trend leads to the fact that until not too long ago (until the year 2002) the world still stored more information in analog than in digital format. Our estimates determine the year 2002 as the “beginning of the digital age” (over 50% digital).

It is useful to put these mind-boggling numbers into context. If we would store the 4.6 optimally compressed zettabytes of 2014 in 730 MB

Information Quantity,

Fig. 1 World's technological capacity to store information 1986–2014 (log on y-axis) (Source: Based on the methodology of Hilbert and López (2011), with own estimates for 2007–2014)



CD-ROM discs (of 1.2 mm thickness), we could build about 20 stacks of discs from the earth to the moon. If we would store the information equivalent in alphanumeric symbols in double-printed books of 125 pages, all the world's landmasses could have been covered with one layer of double-printed book paper back in 1986. By 1993 it would have grown to 6 pages and to 20 pages in the year 2000. By 2007 it would be one layer of books that covers every square centimeter of the world's land masses, two layers by 2010/2011, and some 14 layers by 2014 (letting us literally stand “knee-deep in information”). If we would make piles of these books, we would have about 4500 piles from the Earth to the sun.

Estimating the amount of the world's technological information capacity is only the first step. It can and has been used as input variable to investigate a wide variety of social science questions of the data revolution, including its international distribution, which has shown that the digital divide carries over to the data age (Hilbert 2014b, 2016); the changing nature of content, which has shown that the big data age counts with a larger ratio of alphanumeric text over videos than the pre-2000s (Hilbert 2014c); the crucial role of compression algorithms in the data explosion (Hilbert 2014a), and the impact of data capacity on issues like international trade (Abeliansky and Hilbert 2017).

Further Reading

- Abeliansky, A. L., & Hilbert, M. (2017). Digital technology and international trade: Is it the quantity of subscriptions or the quality of data speed that matters? *Telecommunications Policy*, 41(1), 35–48. <https://doi.org/10.1016/j.telpol.2016.11.001>.
- Aristeas. (200AD, ca). The letter of Aristeas to Philocrates. <http://www.attalus.org/translate/aristeas1.html>.
- Bartlett, J. (1968). William Thompson, Lord Kelvin, popular lectures and addresses [1891–1894]. In *Bartlett's familiar quotations* (14th ed.). Boston: Little Brown & Co.
- Bohn, R., & Short, J. (2009). *How much information? 2009 report on American consumers*. San Diego: Global Information Industry Center of University of California, San Diego.
- Bounie, D. (2003). *The international production and dissemination of information* (Special project on the economics of knowledge Autorità per le Garanzie nelle Comunicazioni). Paris: École Nationale Supérieure des Télécommunications (ENST).
- de S. Pool, I. (1983). Tracking the flow of information. *Science*, 221(4611), 609–613. <https://doi.org/10.1126/science.221.4611.609>.
- de S. Pool, I., Inose, H., Takasaki, N., & Hurwitz, R. (1984). *Communication flows: A census in the United States and Japan*. Amsterdam: North-Holland and University of Tokyo Press.
- Dienes, I. (1986). Magnitudes of the knowledge stocks and information flows in the Hungarian economy. In *Tanulmányok az információgazdaságról* (in Hungarian, pp. 89–101). Budapest.
- Dienes, I. (2010). Twenty figures illustrating the information household of Hungary between 1945 and 2008 (in Hungarian). http://infostat.hu/publikaciok/10_infhazt.pdf.

- Gantz, J., & Reinsel, D. (2012). *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East*. IDC (International Data Corporation) sponsored by EMC.
- Hilbert, M. (2012). How to measure “how much information”? Theoretical, methodological, and statistical challenges for the social sciences. *International Journal of Communication*, 6(Introduction to Special Section on “How to measure ‘How-Much-Information?’”), 1042–1055.
- Hilbert, M. (2014a). How much of the global information and communication explosion is driven by more, and how much by better technology? *Journal of the Association for Information Science and Technology*, 65(4), 856–861. <https://doi.org/10.1002/asi.23031>.
- Hilbert, M. (2014b). Technological information inequality as an incessantly moving target: The redistribution of information and communication capacities between 1986 and 2010. *Journal of the Association for Information Science and Technology*, 65(4), 821–835. <https://doi.org/10.1002/asi.23020>.
- Hilbert, M. (2014c). What is the content of the world’s technologically mediated information and communication capacity: How much text, image, audio, and video? *The Information Society*, 30(2), 127–143. <https://doi.org/10.1080/01972243.2013.873748>.
- Hilbert, M. (2015a). A review of large-scale ‘how much information’ inventories: Variations, achievements and challenges. *Information Research*, 20(4). <http://www.informationr.net/ir/20-4/paper688.html>.
- Hilbert, M. (2015b). *Quantifying the data deluge and the data drought* (SSRN scholarly paper no. ID 2984851). Rochester: Social Science Research Network. <https://papers.ssrn.com/abstract=2984851>.
- Hilbert, M. (2016). The bad news is that the digital access divide is here to stay: Domestically installed bandwidths among 172 countries for 1986–2014. *Telecommunications Policy*, 40(6), 567–581. <https://doi.org/10.1016/j.telpol.2016.01.006>.
- Hilbert, M., & López, P. (2011). The world’s technological capacity to store, communicate, and compute information. *Science*, 332(6025), 60–65. <https://doi.org/10.1126/science.1200970>.
- Hilbert, M., & López, P. (2012a). How to measure the world’s technological capacity to communicate, store and compute information? part I: Results and scope. *International Journal of Communication*, 6, 956–979.
- Hilbert, M., & López, P. (2012b). How to measure the world’s technological capacity to communicate, store and compute information? part II: Measurement unit and conclusions. *International Journal of Communication*, 6, 936–955.
- Ito, Y. (1981). The Johoka Shakai approach to the study of communication in Japan. In C. Wilhoit & H. de Bock (Eds.), *Mass communication review yearbook* (Vol. 2, pp. 671–698). Beverly Hills: Sage.
- Lesk, M. (1997). How much information is there in the world? [lesk.com](http://www.lesk.com/mlesk/ksg97/ksg.html). <http://www.lesk.com/mlesk/ksg97/ksg.html>.
- López, P., & Hilbert, M. (2012). Methodological and statistical background on the world’s technological capacity to store, communicate, and compute information (online document). <http://www.martinhilbert.net/WorldInfoCapacity.html>.
- Lyman, P., Varian, H. R., Dunn, J., Strygin, A., & Swearingen, K. (2000). *How much information 2000*. University of California, at Berkeley.
- Machlup, F. (1962). *The production and distribution of knowledge in the United States*. Princeton: Princeton University Press.
- Porat, M. U. (1977, May). *The information economy: Definition and measurement*. Superintendent of Documents, U.S. Government Printing Office, Washington, DC. 20402 (Stock No. 003-000-00512-7).
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656. <https://doi.org/10.1145/584091.584093>.
- Short, J., Bohn, R., & Baru, C. (2011). *How much information? 2010 report on enterprise server information*. San Diego: Global Information Industry Center at the School of International Relations and Pacific Studies, University of California, San Diego. http://hmi.ucsd.edu/howmuchinfo_research_report_consum_2010.php.
- Turner, V., Gantz, J., Reinsel, D., & Minton, S. (2014). *The digital universe of opportunities: Rich data and the increasing value of the internet of things*. IDC (International Data Corporation) sponsored by EMC.

Information Science

► Informatics

Information Society

Alison N. Novak

Department of Public Relations and Advertising,
Rowan University, Glassboro, NJ, USA

The information age refers to the period of time following the industry growth set forth by the industrial revolution. Although scholars widely debate the start date of this time period, it is often noted that the information age co-occurred with the building and growth in popularity of the Internet. The information age refers to the increasing access, quantification, and collection of digital data, often referred to as big datasets.

Edward Tenner writes that the information age is often called a new age in society because it simultaneously addresses the increasing digital connections between citizens across large distances. Tenner concludes that the information age is largely technology about technology. This suggests that many of the advancements that are connected to the information age are technologies that assist our understanding and connections through other technologies. These include the expansion of the World Wide Web, mobile phones, and GPS devices. The expansion in these technologies has facilitated the ability to connect digitally, collect data, and analyze larger societal trends.

Similarly, the collection and analysis of big datasets was facilitated by many of these information age technologies. The Internet, social networking sites, and GPS systems allow researchers, industry professionals, and government agencies to seamlessly collect data from users to later be analyzed and interpreted. The information age, through the popularization and development of many of these technologies, ushered in a new age of big data research.

Big data in the information age took shape through large, often quantifiable groups of information about individual users, groups, and organizations. As users input data into the Information age technologies, these platforms collected and stored the data for later use. Because the information age elevated the importance and societal value of being digitally connected, users entered large amounts of personal data into these technologies in exchange for digital presence.

John Pavolotsky notes it is for this reason that privacy rose as a central issue in the information age. As users provided data to these technology platforms, legal and ethical issues over who owns the data, who has the right to sell or use the data, and what rights to privacy do users have became critical. It is for this reason that further technologies (such as secure networks) needed to be developed to encourage safety among big data platforms.

As Pavolotsky evidences, the information age is more than just a period in time, it also reshaped values, priorities, and the legal structure of global society. Being connected digitally encouraged more people to purchase personal technologies such as laptops and phones to participate. Further, this change in values similarly altered the demand for high-speed information. Because digital technologies during this period of time encouraged more connections between individuals in the network, information such as current events and trends spread faster than before. This is why the information age is alternatively called a networked society.

Morris and Shin add that the information age changed the public's orientation toward publicly sharing information with a large, diverse, and unknown audience. While concerns of privacy grew during the information age, so did the ability to share and document previously private thoughts, behaviors, and texts. This was not just typical of users but also of media institutions and news organizations. What is and is not considered public information became challenged in an era when previously hidden actions were now freely documented and shared through new technologies such as social networking sites. The effect this user and institutional sharing has had on mass society is still heavily debated. However, this did mean that new behaviors previously not shared or documented in datasets were now freely available to those archiving big datasets and analyzing these technologies.

The information age is also centrally related to changes in global economy, jobs, and development industries. Croissant, Rhoades, and Slaughter suggest that the changes occurring during the information age encouraged learning instructions to focus students toward careers in science, technology, engineering, and mathematics (popularly known as STEM). This focus was because of the rapid expansion in technology and the creation of many new companies and organizations dedicated to expanding the digital commercial front. These new organizations were termed Web 1.0 companies because

of their focus on turning the new values of the information age into valuable commodities. Many of these companies used big datasets collected from user-generated information to target their campaigns and create personalized advertising.

In addition, the information age also affected the structure of banking, financial exchanges, and the global market. As companies expanded their reach using new digital technologies, outsourcing and allocating resources to distant regions became a new norm. Because instantaneous communication across large spaces was now possible and encouraged by the shift in public values, it is easy to maintain control of satellite operations abroad.

The shift to an information society is largely related to the technologies that facilitated big dataset collection and analysis. Although the exact dates of the information society are still debated, the proliferation of social media sites and other digital spaces supports that the information age is ongoing, thus continuing to support the emergence and advancements of big data research.

Cross-References

- ▶ [Mobile Analytics](#)
- ▶ [Network Analytics](#)
- ▶ [Network Data](#)
- ▶ [Privacy](#)
- ▶ [Social Media](#)

Further Reading

- Croissant, J. L., Rhoades, G., & Slaughter, S. (2001). Universities in the information age: Changing work, information, and values in academic science and engineering. *Bulletin of Science Technology Society*, 21(1), 108–118.
- Morris, S., & Shin, H. S. (2002). Social value of public information. *American Economic Review*, 92(5), 1521–1534.
- Pavolotsky, J. (2013). Privacy in the age of big data. *The Business Lawyer*, 69(1), 217–225.

Tenner, E. (1992). Information age at the National Museum of American History. *Technology and Culture*, 33(4), 780–787.

Information Studies

- ▶ [Informatics](#)

Information Systems

- ▶ [Informatics](#)

Information Theory

- ▶ [Informatics](#)

Information Visualisation

- ▶ [Data Visualization](#)

Information Visualization

- ▶ [Data Visualization](#)
- ▶ [Visualization](#)

Informatique

- ▶ [Informatics](#)

Instrument Board

- ▶ [Dashboard](#)

Integrated Data System

Ting Zhang

Department of Accounting, Finance and
Economics, Merrick School of Business,
University of Baltimore, Baltimore, MD, USA

Definition/Introduction

Integrated Data Systems (IDS) typically link individual level administrative records collected by multiple agencies such as k–12 schools, community colleges, other colleges and universities, departments of labor, justice, human resources, human and health services, police, housing, and community services. The systems can be used for quick knowledge-to-practice development cycle (Actionable Intelligence for Social Policy 2017), case management, program or service monitoring, tracking, and evaluation (National Neighborhood Indicators Partnership 2017), research and policy analysis, strategic planning and performance management, and so on. It can also help evaluate how different programs, services, and policies affect individual persons or individual geographic units. The linkages between different agency records are often made through a common individual personal identification number, a shared case number, or a geographic unit.

Purpose of an IDS

With the rising attraction of big data and the exploding need to share existing data, the need to link already collected various administrative records rises. The systems allow government agencies to integrate various databases and bridge the gaps that have traditionally formed within individual agency databases; it can be used for quick knowledge-to-practice development cycle to better address the often interconnected citizens' needs efficiently and effectively (Actionable Intelligence for Social Policy 2017), for case management (National Neighborhood Indicators Partnership 2017), program or service

monitoring, tracking, and evaluation, developing and testing an intervention and monitoring the outcomes (Davis et al. 2014), research and policy analysis, strategic planning and performance management, and so on. It can test social policy innovations through high-speed, low-cost randomized control trials and quasi-experimental approaches, can be used for continuous quality improvement efforts and benefit cost analysis, and can also help provide a complete account of how different programs, services, and policies affect individual persons or individual geographic units to more efficiently and effectively address the often interconnected needs of the citizens (Actionable Intelligence for Social Policy 2017).

Key Elements to Build an IDS

According to Davis et al. (2014) and Zhang and Stevens (2012), typical crucial factors related to a successful IDS include:

- A broad and steady institutional commitment to administrate the system
- Individual-level data (no matter individual persons or individual geographic units) to measure outcomes
- The necessary data infrastructure
- Linkable data fields, such as Social Security numbers, business identifiers, shared case number, and addresses
- The capacity to match various administrative records
- A favorable state interpretation of the data privacy requirements, consistent with federal regulations
- The funding, knowhow, and analytical capacity to work with and maintain the data
- Successfully obtaining participation from multiple data providing agencies with clearance to use those data.

Maintenance

Administrative data records are typically collected by public and private agencies. An IDS

often requires to extract, transform, clean, and link information from various source administrative databases and load it into a data warehouse. Many data warehouses offer a tightly coupled architecture that it usually takes little time to resolve queries and extract information (Widom 1995).

Challenges

Identity Management and Data Quality

One challenge to build an IDS is to have effective and appropriate individual record identity management diagnostics that include consideration of the consequences of gaps in common identifier availability and accuracy. This is the first key step for data quality of IDS information. However, some of the relevant databases, particularly student records, do not include a universally linkable personal identifier, that is, a Social Security number; some databases are unable to ensure that a known to be valid Social Security number is paired with one individual, and only that individual, consistently over time; and some databases are unable to ensure that each individual is associated with only one Social Security number over time (Zhang and Stevens 2012). Zhang and Stevens (2012) included ongoing collection of case studies documenting how SSNs can be extracted, validated, and securely stored offline. With the established algorithms required for electronic financial transactions, spreading adoption of electronic medical records and rising interest in big data, there is an extensive, and rapidly growing, literature illustrating probabilistic matching solutions and various software designs to address the identity management challenge. Often the required accuracy threshold is application specific; assurance of an exact match may not be required for some anticipated longitudinal data system uses (Zhang and Stevens 2012).

Data Privacy

To build and use an IDS, issues related to privacy of personal information within the system is important. Many government agencies have relevant regulations. For example, a nationally wide-

known law is the Family Educational Rights and Privacy Act (FERPA) that defines when student information can be disclosed and data privacy practices (U.S. Department of Education 2017). Similarly Health Insurance Portability and Accountability Act of 1996 (HIPAA) addresses the use and disclosure of health information (U.S. Department of Health & Human Services 2017).

Ethics

Most IDS taps individual person's information. When using IDS information, in order not to misuse personal information, extra caution is needed. Institutional review boards are often needed when conducting research involving human subjects.

Data Sharing

To build an IDS, a favorable state interpretation of the data privacy requirements, consistent with federal regulations and clearance to use the data for the IDS, is critical. For example, some state education agencies have been reluctant to share their education records, largely due to narrow state interpretations of the confidentiality provisions of FERPA and its implementing regulations (Davis et al. 2014). Corresponding data sharing agreements need to be in place.

Data Security

During the process of building, transferring, maintaining, and using IDS information, the data security issue in an IDS center is particularly important. Measures to ensure data security and information privacy and confidentiality becomes the key factors for an IDS' vigor and sustainability. Fortunately, many of the US current IDS centers have had experience maintaining confidential administrative records for years or even decades. However, facing the convenience of web access to maintain the continued data security and sustainability often requires updated data protection techniques. The federal, state, and local government has important roles in safeguarding data and data use.

Examples

Example of IDS in the United States include:

Chapin Hall's Planning for Human Service Reform Using Integrated Administrative Data

Jacob France Institute's database for education, employment, human resources, and human services

Juvenile Justice and Child Welfare Data Cross-over Youth Multi-Site Research Study

Actionable Intelligence for Social Policy's integrated data systems initiatives for policy analysis and program reform

Florida's Common Education Data Standards (CEDS) Workforce Workgroup and the later Florida Education & Training Placement Information Program

Louisiana Workforce Longitudinal Data System (WLDS) housed at the Louisiana Workforce Commission

Minnesota's iSEEK data, managed by an organization called iSEEK Solutions

Heldrich Center data at Rutgers University

Ohio State University's workforce longitudinal administrative database

University of Texas Ray Marshall Center database

Virginia Longitudinal Data System

Washington's Career Bridge, managed by the Workforce Training and Education Coordinating Board

Connecticut's Preschool through Twenty and Workforce Information Network (P-20 WIN)

Delaware Department of Education's Education Insight Dashboard

Georgia Department of Education's Statewide Longitudinal Data System and Georgia's Academic and Workforce Analysis and Research Data System (GA AWARDS)

Illinois Longitudinal Data System

Indiana Network of Knowledge (INK)

Maryland Longitudinal Data System

Missouri Comprehensive Data System

Ohio Longitudinal Data Archive (OLDA)

South Carolina Longitudinal Information Center for Education (SLICE)

Texas Public Education Information Resource (TPEIR) and Texas Education Research Center (ERC)

Washington P-20W Statewide Longitudinal Data System

Conclusion

Integrated Data Systems (IDS) typically link individual level administrative records collected by multiple agencies. The systems can be used for case management, program or service monitoring, tracking, and evaluation, research and policy analysis, etc. A successful IDS often requires a broad and steady institutional commitment to administer the system, individual-level data, the necessary data infrastructure, linkable data fields, capacity and knowhow to match various administrative records and maintain it, data access permission, and data privacy procedures. Main challenges to build a sustainable IDS include identity management, data quality, data privacy, ethics, data sharing, and data security. There are many IDS in the United States.

Further Reading

- Actionable Intelligence for Social Policy. (2017). Integrated Data Systems (IDS). Retrieved in March 2017 from the World Wide Web at <https://www.aisp.upenn.edu/integrated-data-systems/>.
- Davis, S., Jacobson, L., & Wandner, S. (2014). *Using workforce data quality initiative databases to develop and improve consumer report card systems*. Washington, DC: Impaq International.
- National Neighborhood Indicators Partnership. (2017). Resources on Integrated Data Systems (IDS). Retrieved in March 2017 from the World Wide Web at <http://www.neighborhoodindicators.org/resources-integrated-data-systems-ids>.
- U.S. Department of Education. (2017). Family Educational Rights and Privacy Act (FERPA). Retrieved on May 14, 2017 from the World Wide Web <https://ed.gov/policy/gen/guid/fpco/ferpa/index.html>.
- U.S. Department of Health & Human Services. (2017). Summary of the HIPAA Security Rule. Retrieved on May 14, 2017 from the World Wide Web <https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/>.
- Widom, J. (1995). "Research problems in data warehousing." *CIKM '95 Proceedings of the fourth international conference on information and knowledge management* (pp. 25–30). Baltimore.
- Zhang, T., & Stevens, D. (2012). Integrated data system person identification: Accuracy requirements and methods. Jacob France Institute. Available at SSRN: <https://ssrn.com/abstract=2512590> or <https://doi.org/10.2139/ssrn.2512590> and <http://www.workforcedqc.org/sites/default/files/images/JFI%20wdqi%20research%20report%20January%202014.pdf>.

Intelligent Agents

► Artificial Intelligence

Intelligent Transportation Systems (ITS)

Laurie A. Schintler
George Mason University, Fairfax, VA, USA

Overview

In the last half-century, digital technologies have transformed the surface transportation sector – that is, highway, rail, and public transport. It is in this context that the concept of Intelligent Transportation Systems (ITS) transpired. ITS generally pertains to the use of advanced technologies and real-time information for monitoring, managing, and controlling surface transportation modes, services, and systems. The first generation of ITS, referred to as Intelligent Vehicle Highway Systems (IVHS), was focused primarily on applying Information and Communications Technology

(e.g., computers, robotics, and control software) to highway systems for improving mobility, safety, and productivity. With new and expanding sources of big data, coupled with advancements in analytical and computational capabilities and capacities, we are on the cusp of another technological revolution in surface transportation. This latest phase of ITS is more sophisticated, integrated, and broader in scope and purpose than before. However, while state-of-the-art ITS applications promise to benefit society in many ways, they also come with various technical, institutional, ethical, legal, and informational issues, challenges, and complexities.

Opportunities, Prospects, and Applications

Intelligent Transportation Systems (ITS) support various functions and activities (see Table 1). Different types and sources of big data, along with big data analytics, help support the goals and objectives of each of these systems and applications.

Novel sources of big data create fresh opportunities for the management, control, and provision of transportation and mobility. First, the

Intelligent Transportation Systems (ITS), Table 1 ITS uses and applications

Application	Aims and objectives
Traffic and travel information	To provide continuous and reliable traffic and travel data and information for transportation producers and consumers
Traffic and public transport management	To improve traffic management in cities and regions for intelligent traffic signal control, incident detection and management, lane control, speed limits enforcement, etc.
Navigation services	To provide route guidance to transportation users
Smart ticketing and pricing	To administer and collect transportation fees for the pricing of transport services, based on some congestion, emissions, or some other consideration, and to facilitate “smart ticketing” systems
Safety and security	To reduce the number and severity of accidents and other safety issues
Freight transport and logistics	To gather, store, analyze, and provide access to cargo data for helping freight operators to make better decisions
Intelligent mobility and co-modality services	To provide real-time information and analysis to transportation users for facilitating trip planning and management
Transportation automation (smart and connected vehicles)	To enable fully or partially automated movement of vehicles or fleets of vehicles

Source: Adapted from Giannopoulos et al. (2012)

proliferation of GPS-equipped (location-enabled) devices, such as mobile phones, RFID tags, and smart cards, enable real-time and geographically precise tracking of people, information, animals, and goods. Second, data produced by crowdsourcing platforms, Web 2.0 “apps,” and social media – actively and passively – help to facilitate an understanding of the transportation needs, preferences, and attitudes of individuals, organizations, firms, and communities. Third, satellites, drones, and other aerial sensors offer an ongoing and detailed view of our natural and built environment, enabling us to better understand the factors that not only affect transportation (e.g., weather conditions) but that are also affected by transportation (e.g., land use patterns, pollution). Fourth, the Internet of Things (IoT), which comprises billions of interconnected sensors tied to the Internet, combined with Cyber-Physical Systems, are monitoring various aspects and elements of transportation systems and operations for anomaly detection and system control. Fifth, transportation automation, which comes in various forms – from Automated vehicles (AVs) to drones and droids – is producing vast amounts of streaming data on traffic and road conditions and other aspects of the environment. Lastly, video cameras and other modes of surveillance, which have become ubiquitous, are contributing to a massive collection of dynamic, unstructured data, which provide new resources for monitoring transportation systems and their use.

Big data analytics, tools, and techniques also are playing a vital role in ITS, particularly for mining and analyzing big data to understand and anticipate issues and problems (e.g., accidents, bottlenecks) and ultimately to develop the intelligence needed to act and intervene efficiently and appropriately to enhance transportation systems. Deep neural learning – a powerful form of machine learning that mimics aspects of information processing in the human brain – is running behind the scenes literally everywhere to optimize and control transportation modes, systems, and services. Specifically, deep learning is being used for transportation performance evaluation, traffic and congestion prediction, avoidance of incidents and accidents, vehicle identification,

traffic signal optimization, ridesharing, public transport, visual recognition tasks, among others. New developments and breakthroughs in Natural Language Processing (NLP) (including sentiment analysis) and image, video, and audio processing facilitate the analysis and mining of unstructured big data, such as that produced by social media feeds, news stories, and video surveillance cameras. Innovations in network analysis and visualization tools and algorithms, along with improvements in computational and storage capacity, now enable large, complex, and dynamic networks (e.g., logistics, critical infrastructure) to be tracked and analyzed in real-time. Finally, cloud robotics, which combines cloud computing and machine learning (e.g., reinforcement learning), is the “brain” behind automated systems – for example, autonomous vehicles, enabling them to learn from their environment and from each other to adapt and respond in an optimal way.

In ITS, technologies are also crucial for facilitating the storage, communication, and dissemination of data and information within and across organizations and to travelers and other transportation consumers. New and emerging technological systems and platforms are leading to various innovations in this regard. For example, methods for transmitting data in both public and commercial settings have evolved from wired systems to wireless networks supported by cloud platforms. Modes of disseminating messages to the public (e.g., advisory statements) have shifted from static traffic signage and radio and television broadcasting to intelligent Variable Message Signs (VMS), mobile applications, and in-vehicle information services. Blockchain technology is just beginning to replace traditional database systems, particularly for vetting, archiving, securing, and sharing information on transportation transactions and activities, such as those tied to logistics and supply chains and mobility services.

Issues, Challenges, and Complexities

While big data (and big data analytics) bring many benefits for transportation and mobility, their use in this context also raises an array of ethical, legal,

and social downsides and dangers. One serious issue, in particular, is algorithmic bias and discrimination, a problem in which the outcomes of machine learning models and decisions based on them have the potential to disadvantage and even harm certain segments of the population. One source of this problem is the data used for training, testing, and validating such models. Algorithms learn from the real-world; accordingly, if there are societal gaps and disparities reflected in the data in the first place, then the output of machine learning and its application and use for decision-making may reinforce or even perpetuate inequalities and inequities. This problem is particularly relevant for ITS, as people, organizations, and places are not affected or benefited by transportation in the same way. For example, people of color, indigenous populations, women, and the poor generally have lower mobility levels and fewer transportation options than others. Some of these same groups and communities are disproportionately negatively impacted by transportation externalities, such as noise, air, and water pollution. Privacy is another matter. Many of the big data sources and enabling technologies used in Intelligent Transportation Systems contain sensitive information on the activities of individuals and companies in time and space. For instance, ITS applications that rely on photo enforcement have the potential for privacy infringement, given their active role in tracking and identifying vehicles.

Another set of related concerns stem from the increasing presence and involvement of technology companies in designing, implementing, and managing ITS in public spaces. The private sector's goals and objectives tend to be incongruent with those of the public sector, where the former is generally interested in maximizing profit and rate-of-return and the latter enhancing societal welfare. Accordingly, the collection and use of big data, along with big data algorithms, has the potential to reflect the interests and motivations – and values – of companies rather than those of the public-at-large. This situation also compounds knowledge and power asymmetries and imbalances between the public and private sectors, where information on transportation systems in the public sphere are increasingly in the

hands of commercial entities, rather than planners and government managers.

The use of big data in ITS applications raises various technical and informational challenges. Ensuring the interoperability of vehicles, mobile devices, infrastructure, operations centers, and other ITS elements pose significant challenges, particularly given that new technologies and information systems comprise many moving parts that require careful integration and coupling in real-time. Other challenges relate to assessing how accurately big data captures different aspects of transportation systems and integrating big data with conventional data sources (e.g., traffic counts and census records).

Conclusion

Nearly every aspect of our lives depends critically on transportation and mobility. Transportation systems are vital for the production, consumption, distribution, and exchange of goods and services and accordingly, are critical drivers of economic growth, development, and prosperity. Moreover, as a means for accessing opportunities and activities, such as healthcare, shopping, and entertainment, transportation is a social determinant of health and well-being. In these regards, new and advancing ITS applications, supported by big data, big data analytics, and emerging technologies, help maximize the full potential of surface transportation. At the same time, policies, standards, and practices, including ethical and legal frameworks, are needed to ensure that the benefits of ITS are equitably distributed within and across communities and that no one is disproportionately negatively impacted by transportation innovation.

Cross-References

- ▶ [Cell Phone Data](#)
- ▶ [Mobile Analytics](#)
- ▶ [Sensor Technologies](#)
- ▶ [Smart Cities](#)
- ▶ [Supply Chain and Big Data](#)
- ▶ [Transportation Visualization](#)

Further Reading

- Chen, Z., & Schintler, L. A. (2015). Sensitivity of location-sharing services data: evidence from American travel pattern. *Transportation*, 42(4), 669–682.
- Fries, R. N., Gahrooei, M. R., Chowdhury, M., & Conway, A. J. (2012). Meeting privacy challenges while advancing intelligent transportation systems. *Transportation Research Part C: Emerging Technologies*, 25, 34–45.
- Giannopoulos, G., Mitsakis, E., Salanova, J. M., Dilara, P., Bonnel, P., & Punzo, V. (2012). Overview of Intelligent Transport Systems (ITS) developments in and across transport modes. *JRC Scientific and Policy Reports*, 1–34.
- Haghighat, A. K., Ravichandra-Mouli, V., Chakraborty, P., Esfandiari, Y., Arabi, S., & Sharma, A. (2020). Applications of deep learning in intelligent transportation systems. *Journal of Big Data Analytics in Transportation*, 2(2), 115–145.
- Schintler, L. A., & McNeely, C. L. (2020). Mobilizing a culture of health in the era of smart transportation and automation. *World Medical & Health Policy*, 12(2), 137–162.
- Sumalee, A., & Ho, H. W. (2018). Smarter and more connected: Future intelligent transportation system. *IATSS Research*, 42(2), 67–71.
- Zhang, J., Wang, F. Y., Wang, K., Lin, W. H., Xu, X., & Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4), 1624–1639.

Interactive Data Visualization

Andreas Veglis

School of Journalism and Mass Communication,
Aristotle University of Thessaloniki,
Thessaloniki, Greece

Definition

Data visualization is a modern branch of descriptive statistics that involves the creation and study of the visual representation of data. It is the graphical display of abstract information for data analysis and communication purposes. Static data visualization offers only precomposed “views” of data. Interactive data visualization supports multiple static views in order to present a variety of perspectives on the same information. Important stories include “hidden” data, and interactive

data visualization is the appropriate mean to discover, understand, and present these stories. In interactive data visualization there is a user input (a control of some aspect of the visual representation of information), and the changes made by the user must be incorporated into the visualization in a timely manner. They are based on existing sets of data, and obviously this subject is strongly related with the issue of big data. Data visualizations is the best method in order to transform chunks of data to meaningful information (Ward et al. 2015).

History

Although people have been using tables in order to arrange data since the second century BC, the idea of representing quantitative information graphically first appeared in the seventeenth century. Rene Descartes, who was a French philosopher and mathematician, proposed a two-dimensional coordinate system for displaying values, consisting of a horizontal axis for one variable and a vertical axis for another, primarily as a graphical means of performing mathematical operations. In the eighteenth century William Playfair began to exploit the potential of graphics for the communication of quantitative, by developing many of the graphs that are commonly used today. He was the first to employ a line moving up and down as it progressed from left to right to show how values changed through time. He invented the bar graph, as well as the pie chart. In the 1960s Jacques Bertin proposed that visual perception operates according to rules that can be followed to express information visually in ways that represented it intuitively, clearly, accurately, and efficiently. Also John Tukey, a statistics professor set the basis of the exploratory data analysis, by demonstrating the power of data visualization as a means for exploring and making sense of quantitative data (Few 2013).

In 1983, Edward Tufte published his groundbreaking book “The Visual Display of Quantitative Information,” in which he distinguished between the effective ways of displaying data visually and the ways that most people are doing

it without much success. Also around this time, William Cleveland extended and refined data visualization techniques for statisticians. At the end of the century, the term information visualization was proposed. In 1999, Stuart Card, Jock Mackinlay, and Ben Shneiderman published their book entitled “Readings in Information Visualization: Using Vision to Think.” Moving to the twenty-first century, Colin Ware published two books entitled “Information Visualization: Perception for Design (2004) and Visual Thinking for Design (2008)” in which he compiled, organized, and explained what we have learned from several scientific disciplines about visual thinking and cognition and applied that knowledge to data visualization (Few 2013).

Since the turn of the twenty-first century, data visualization has been popularized, and it has reached the masses through commercial software products that are distributed through the web. Many of these data visualization products promote more superficially appealing esthetics and neglect the useful and effective data exploration, sense-making, and communication. Nevertheless there are a few serious contenders that offer products which help users fulfill data visualization potential in practical and powerful ways.

From Static to Interactive

Visualization can be categorized into static and interactive. In the case of the static visualization, there is only one view of data, and in many occasions, multiple cases are needed in order to fully understand the available information. Also the number of dimensions of data is limited. Thus representing multidimensional datasets fairly in static images is almost impossible. Static visualization is ideal when alternate views are neither needed nor desired and is special suited for static medium (e.g., print) (Knaflitz 2015). It is worth mentioning that infographics are also part of the static visualization. Infographics (or information graphics) are graphic visual representations of data or knowledge, which are able to present complex information quickly and clearly. Infographics are being used for many years, and

recently the availability of many easy-to-use free tools have made the creation of infographics available to every Internet user (Murray 2013).

Of course static visualizations can also be published on the World Wide Web in order to disseminate more easily and rapidly. Publishing on the web is considered to be the quickest way to reach a global audience. An online visualization is accessible by any Internet user that employs a recent web browser, regardless of the operating system (Windows, Mac, Linux, etc.) and device type (laptop, desktop, smartphone, tablet). But the true capabilities of the web are being exploited in the case of interactive data visualization.

Dynamic, interactive visualizations can empower people to explore data on their own. The basic functions of most interactive visualization tools have been set back in 1996, when Ben Shneiderman proposed a “Visual Information-Seeking Mantra” (overview first, zoom and filter, and then details on demand). The above functions allow data to be accessible from every user, from the one who is just browsing or exploring the dataset to the one who approaches the visualization with a specific question in mind. This design pattern is the basic guide for every interactive visualization today.

An interactive visualization should initially offer an overview of the data, but it must also include tools for discovering details. Thus it will be able to facilitate different audiences, from those who are new to the subject to those who are already deeply familiar with the data. Interactive visualization may also include animated transitions and well-crafted interfaces in order to engage the audience to the subject it covers.

User Control

In the case of interactive data visualization, users interact with the visualization by introducing a number of input types. Users can zoom in a particular part of an existing visualization, pinpoint an area that interest them, select an option from an offered list, choose a path, and input numbers or text that customize the visualization. All the previous mentioned input types can be accomplished

by using keyboard, mice, touch screen, and other more specialized input devices. With the help of these input actions, users can control both the information being represented on the graph or the way that the information is being presented. In the second case, the visualization is usually part of a feedback loop. In most cases the actual information remains the same, but the representation of the information does change. One other important parameter in the interactive data visualizations is the time it takes for the visualization to be updated after the user has introduced an input. A delay of more than 20 ms is noticeable by most people. The problem is that when large amounts of data are involved, this immediate rendering is impossible.

Interactive framerate is a term that is often being used to measure the frequency with which a visualization system generates an image. In case that the rapid response time, which is required for interactive visualization, is not feasible, there are several approaches that have been explored in order to provide people with rapid visual feedback based on their input. These approaches include:

Parallel rendering: in this case the image is being rendered simultaneously by two or more computers (or video cards). Different frames are being rendered at the same time by different computers, and the results are transferred over the network for display on the user's computer.

Progressive rendering: in this case a framerate is guaranteed by rendering some subset of the information to be presented. It also provides progressive improvements to the rendering when the visualization is no longer changing.

Level-of-detail (LOD) rendering: in this case simplified representations of information are rendered in order to achieve the desired frame rate, while a user is providing input. When the user has finished manipulating the visualization, then the full representation is used in order to generate a still image.

Frameless rendering: in this type of rendering, the visualization is not presented as a time series of images. Instead a single image is

generated where different regions are updated over time.

Types of Interactive Data Visualizations

The information and more specifically statistical information is abstract, since it describes things that are not physical. It can concern education, sales, diseases, and various other things. But everything can be displayed visually, if the way is found to give them a suitable form. The transformation of the abstract into physical representation can only succeed if we understand a bit about visual perception and cognition. In other words, in order to visualize data effectively, one must follow design principles that are derived from an understanding of human perception.

Heer, Bostock and Ogievetsky (2010) defined the types (and also their subcategories) of data visualization:

- (i) Time series data (index charts, stacked graphs, small multiples, horizon graphs)
- (ii) Statistical distributions (stem-and-leaf plots, Q-Q plots, scatter plot matrix (SPLOM), parallel coordinates)
- (iii) Maps (flow maps, choropleth maps, graduated symbol maps, cartograms)
- (iv) Hierarchies (node-link diagrams, adjacency diagrams, enclosure diagrams)
- (v) Networks (force-directed layout, arc diagrams, matrix views)

Tools

There are a lot of tools that can be used for creating interactive data visualizations. All of them are either free or offer a free version (except a paid version that includes more features). According to datavisualization.ch, the list of the tools that most users employ includes: Arbor.js, CartoDB, Chroma.js, Circos, Cola.js, ColorBrewer, Cubism.js, Cytoscape, D3.js, Dance.js, Data.js, DataWrangler, Degrafa, Envision.js, Flare, GeoCommons, Gephi, Google Chart Tools, Google Fusion Tables, I Want

Hue, JavaScript InfoVis Toolkit, Kartograph, Leaflet, Many Eyes, MapBox, Miso, Modest Maps, Mr. Data Converter, Mr. Nester, NVD3.js, NodeBox, OpenRefine, Paper.js, Peity, Polymaps, Prefuse, Processing, Processing.js, Protovis, Quadrigram, R, Raphael, Raw, Recline.js, Rickshaw, SVG Crowbar, Sigma.js, Tableau Public, Tabula, Tangle, Timeline.js, Unfolding, Vega, Visage, and ZingCharts.

Conclusion

Data visualization is a significant discipline that is expected to become even more important as we gradually moving, as a society, in the era of big data. Especially the case of interactive data visualization allows data analysts to convey complex data to meaningful information that can be searched, explored, and understood by end users.

Cross-References

- [Business Intelligence](#)
- [Tableau Software](#)
- [Visualization](#)

Further Reading

- Few, S. (2013). Data visualization for human perception. In S. Mads & D. R. Friis (Eds.), *The encyclopedia of human-computer interaction* (2nd ed.). Aarhus: The Interaction Design Foundation. <http://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception>. Accessed 12 July 2016.
- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM*, 53(6), 59–67.
- Knaff, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Hoboken, New Jersey: John Wiley & Sons Inc.
- Murray, S. (2013). *Interactive data visualization for the web*. Sebastopol, CA: O'Reilly Media, Inc.
- Ward, M., Grinstein, G., & Keim, D. (2015). *Interactive data visualization: Foundations, techniques, and applications*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

International Development

Jon Schmid

Georgia Institute of Technology, Atlanta, GA, USA

Big data can affect international development in two primary ways. First, big data can enhance our understanding of underdevelopment by expanding the evidence base available to researchers, donors, and governments. Second, big data-enabled applications can affect international development directly by facilitating economic behavior, monitoring local conditions, and improving governance. The following sections will look first at the role of big data in increasing our understanding of international development and then look at examples where big data has been used to improve the lives of the world's poor.

Big Data in International Development Research

Data quality and data availability tend to be low in developing countries. In Kenya, for example, poverty data was last collected in 2005, and income surveys in other parts of sub-Saharan Africa often take up to 3 years to be tabulated. When national income-accounting methodologies were updated in Ghana (2010) and Nigeria (2014), GDP calculations had to be revised upward by 63% and 89%, respectively. Poor-quality or stale data prevent national policy makers and donors from making informed policy decisions.

Big data analytics has the potential to ameliorate this problem by providing alternative methods for collecting data. For example, big data applications may provide a novel means by which national economic statistics are calculated. The Billion Prices Project – started by researchers at the Massachusetts Institute of Technology – uses daily price data from hundreds of online retailers to calculate changes in price levels. In

countries where inflation data is unavailable – or in cases such as Argentina where official data is unreliable – these data offer a way of calculating national statistics that does not require a high-quality national statistics agency.

Data from mobile devices is a particularly rich source of data in the developing world. Roughly 20% of mobile subscriptions are held by individuals that earn less than 5 \$ a day. Besides emitting geospatial, call, and SMS data, mobile devices are increasingly being used in the developing world to perform a broad array of economic functions such as banking and making purchases. In many African countries (nine in 2014), more people have online mobile money accounts than have traditional bank accounts. Mobile money services such as M-Pesa and MTN Money produce trace data and thus offer intriguing possibilities for increasing understanding of spending and saving behavior in the developing world. As the functionality provided by mobile money services extends into loans, money transfers from abroad, cash withdrawal, and the purchase of goods, the data yielded by these platforms will become even richer.

The data produced by mobile devices has already been used to glean insights into complex economic or social systems in the developing world. In many cases, the insights into local economic conditions that result from the analysis of mobile device data can be produced more quickly than national statistics. For example, in Indonesia the UN Global Pulse monitored tweets about the price of rice and found them to be highly correlated with national spikes in food prices. The same study found that tweets could be used to identify trends in other types of economic behavior such as borrowing. Similarly, research by Nathan Eagle has shown that reductions in additional airtime purchases are associated with falls in income. Researchers Han Wang and Liam Kilmartin examined Call Detail Record (CDR) data generated from mobile devices in Uganda and identified differences in the way that wealthy and poor individuals respond to price discounts. The researchers also used the data to identify centers of economic activity within Uganda.

Besides providing insight into how individuals respond to price changes, big data analytics allows researchers to explore the complex ways in which the economic lives of the poor are organized. Researchers at Harvard's Engineering Social Systems lab have used mobile phone data to explore the behavior of inhabitants of slums in Kenya. In particular, the authors tested theories of rural-to-urban migration against spatial data emitted by mobile devices. Some of the same researchers have used mobile data to examine the role of social networks on economic development and found that diversity in individuals' network relationships is associated with greater economic development. Such research supports the contention that insular networks – i.e., highly clustered networks with few ties to outside nodes – may limit the economic opportunities that are available to members.

Big data analytics are also being used to enhance understanding of international development assistance. In 2009, the College of William and Mary, Brigham Young University, and Development Gateway created AidData (aiddata.org), a website that aggregates data on development projects to facilitate project coordination and provide researchers with a centralized source for development data. AidData also maps development projects geospatially and links donor-funded projects to feedback from the project's beneficiaries.

Big Data in Practice

Besides expanding the evidence base available to international development scholars and practitioners, large data sets and big data analytic techniques have played a direct role in promoting international development. Here the term “development” is considered in its broad sense as referring not to a mere increase in income, but to improvements in variables such as health and governance.

The impact of infectious diseases on developing countries can be devastating. Besides the obvious humanitarian toll of outbreaks, infectious diseases prevent the accumulation of human capital and strain local resources. Thus there is great

potential for big data-enabled applications to enhance epidemiological understanding, mitigate transmission, and allow for geographically targeted relief. Indeed, it is in the tracking of health outcomes that the utility of big data analytics in the developing world has been most obvious. For example, Amy Wesolowski and colleagues used mobile phone data from 15 million individuals in Kenya to understand the relationship between human movement and malaria transmission. Similarly, after noting in 2008 that search trends could be used to track flu outbreaks, researchers at Google.org have used data on searches for symptoms to predict outbreaks of the dengue virus in Brazil, Indonesia, and India. In Haiti, researchers from Columbia University and the Karolinska Institute used SIM card data to track the dispersal of people following a cholera outbreak. Finally, the Centers for Disease Control and Prevention used mobile phone data to direct resources to appropriate areas during the 2014 Ebola outbreak.

Big data applications may also prove useful in improving and monitoring aspects of governance in developing countries. In Kenya, India, and Pakistan, witnesses of public corruption can report the incident online or via text message to a service called “I Paid A Bribe.” The provincial government in Punjab, Pakistan, has created a citizens’ feedback model, whereby citizens are solicited for feedback regarding the quality of government services they received via automated calls and texts. In effort to discourage absenteeism in India and Pakistan, certain government officials are provided with cell phones and required to text geocoded pictures of themselves at jobsites. These mobile government initiatives have created a rich source of data that can be used to improve government service delivery, reduce corruption, and more efficiently allocate resources.

Applications that exploit data from social media have also proved useful in monitoring elections in sub-Saharan Africa. For example, Aggie, a social media tracking software designed to monitor elections, has been used to monitor elections in Liberia (2011), Ghana (2012), Kenya (2013), and Nigeria (2011 and 2014). The Aggie system is first fed with a list of predetermined keywords,

which are established by local subject matter experts. The software then crawls social media feeds – Twitter, Facebook, Google+, Ushahidi, and RSS – and generates real-time trend visualizations based on keyword matches. The reports are monitored by a local Social Media Tracking Center, which identifies instances of violence or election irregularities. Flagged incidents are passed on to members of the election commission, police, or other relevant stakeholders.

The history of international economic development initiatives is fraught with would-be panaceas that failed to deliver. White elephants – large-scale capital investment projects for which the social surplus is negative – are strewn across poor countries as reminders of the preferred development strategies of the past. While more recent approaches to reducing poverty that have focused on improving institutions and governance within poor countries may produce positive development effects, the history of development policy suggests that optimism should be tempered. The same caution holds in regard to the potential role of big data in international economic development. Martin Hilbert’s 2016 systematic review article rigorously enumerates both the causes for optimism and reasons for concern. While big data may assist in understanding the nature of poverty or lead to direct improvements in health or governance outcomes, the availability and ability to process large data sets are not a panacea.

Cross-References

- [Economics](#)
- [Epidemiology](#)
- [International Development](#)
- [World Bank](#)

Further Reading

- Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, 34(1), 135–174.
- Wang, H., & Kilmartin, L. (2014). Comparing rural and urban social and economic behavior in Uganda:

- Insights from mobile voice service usage. *Journal of Urban Technology*, 21(2), 61–89.
- Wesolowski, A., et al. (2012). Quantifying the impact of human mobility on malaria. *Science*, 338(6104), 267–270.
- World Economic Forum. (2012). Big data, big impact: New possibilities for international development. In *Big data, big impact: New possibilities for international development*, Cologny/Geneva, Switzerland: World Economic Forum. http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf.

International Labor Organization

Jennifer Ferreira
Centre for Business in Society, Coventry
University, Coventry, UK

Every day people across the world in both developed and developing economies are creating an ever-growing ocean of digital data. This “big data” represents a new resource for international organizations with the potential to revolutionize the way policies, programs, and projects are generated. The International Labour Organization (ILO) is no exception to this and has begun to discuss and engage with the potential uses of big data to contribute to its agenda.

Focus

The ILO, founded in 1919 in the wake of the First World War, became the first specialized agency of the United Nations. It focuses on labor issues including child labor, collective bargaining, corporate social responsibility, disability, domestic workers, forced labor, gender equality, informal economy, international labor migration, international labor standards, labor inspection, micro-finance, minimum wages, rural development, and youth employment. By 2013 the ILO had 185 members (of the 193 member states of the United Nations). Among its multifarious activities, it is widely known for its creation of

Conventions and Recommendations (189 and 203, respectively by, 2014) related to labor market standards.

Where Conventions are ratified, come into force, and are therefore legally binding, they create a legal obligation for ratifying nations. For many Conventions even in countries where they are not ratified, they are often adopted and interpreted as the international labor standard. There have been many important milestones created by the ILO to shape the landscape to encourage the promotion of improved working lives globally, although a significant milestone is often considered to be the 1998 Declaration on the Fundamental Principles and Rights to Work which had four key components: the right of workers to associate freely and collectively, the end of forced and compulsory labor, the end of child labor, and the end of unfair discrimination among workers. ILO members have an obligation to work toward these objectives and respect the principles which are embedded in the Conventions.

Decent Work Agenda

The ILO believes that work plays a crucial role in the well-being of workers and families and therefore the broader social and economic development of individuals, communities, and societies. While the ILO works on many issues related to employment, their key agenda which has dominated activities in recent decades is “decent work.”

“Decent work” refers to an aspiration for people to have a work that is productive, provides a fair income with security and social protection, safeguards basic rights, and offers equal opportunities and treatment, opportunities for personal development, and a voice in society. “Decent work” is central to efforts to reduce poverty and is a path to achieving equitable, inclusive, and sustainable development; ultimately it is seen as a feature which underpins peace and security in communities and societies (ILO 2014a).

The “decent work” concept was formulated by the ILO in order to identify the key priorities to

focus their efforts. “Decent work” is designed to reflect priorities on the social, economic, and political agenda of countries as well as the international system. In a relatively short time, this concept has formed an international consensus among government, employers, workers, and civil equitable globalization, a path to reduce poverty as well as inclusive and sustainable development. The overall goal of “decent work” is to instigate positive change in/for people at all spatial scales.

Putting the decent work agenda into practice is achieved through the implementation of the ILO’s four strategic objectives, with gender equality as a crosscutting objective:

1. Creating jobs to foster an economy that generates opportunities for investment, entrepreneurship, skills development, job creation, and sustainable livelihoods.
2. Guaranteeing rights at work in order to obtain recognition for work achieved as well as respect for the rights of all workers.
3. Extending social protection to promote both inclusion and productivity of all workers. To be enacted by ensuring both women and men experience safe working conditions, allowing free time, taking into account family and social values and situations, and providing compensation where necessary in the case of lost or reduced income.
4. Promoting social dialogue by involving both workers and employers in the organizations in order to increase productivity, avoid disputes and conflicts at work, and more broadly build cohesive societies.

ILO Data

The ILO produces research on important labor market trends and issues to inform constituents, policy makers, and the public about the realities of employment in today’s modern globalized economy and the issues facing workers and employers in countries at all development stages. In order to do so, it draws on data from a wide variety of sources.

The ILO is a major provider of statistics as these are seen as important tools to monitor progress toward labor standards. In addition to the maintenance of key databases (ILO 2014b) such as LABOURSTA, it also publishes compilations of labor statistics, such as the Key Indicators of Labour Markets (KILM) which is a comprehensive database of country level data for key indicators in the labor market which is used as a research tool for labor market information. Other databases include the ILO STAT, a series of databases with labor-related data; NATLEX which includes legislation related to labor markets, social security, and human rights; and NORMLEX which brings together ILO labor standards and national labor and security laws (ILO 2014c). The ILO database provides a range of datasets with annual labor market statistics including over 100 indicators worldwide including annual indicators as well as short-term indicators, estimates and projections of total population, and labor force participation rates.

Statistics are vital for the development and evaluation of labor policies, as well as more broadly to assess progress toward key ILO objectives. The ILO supports member states in the collection and dissemination of reliable and recent data on labor markets. While the data produced by the ILO are both wide ranging and widely used, they are not considered by most to be “big data,” and this has been recognized.

ILO, Big Data, and the Gender Data

In October 2014, a joint ILO-Data2X roundtable event held in Switzerland identified the importance of developing innovative approaches to the better use of technology to include big data, in particular where it can be sourced and where innovations can be made in survey technology. This event, which brought together representatives from national statistics offices, key international and regional organizations, and nongovernmental organizations, was organized to discuss where there were gender data gaps, particularly focusing on informal and unpaid work as well as agriculture. These discussions

were sparked by wider UN discussions about the data revolution and the importance of development data in the post-2015 development agenda. It is recognized that big data (including administrative data) can be used to strengthen existing collection of gender statistics, but there need to be more efforts to find new and innovative ways to work with new data sources to meet a growing demand for more up to date (and frequently updating) data on gender and employment (United Nations, 2013). The fundamental goal of the discussion was to improve gender data collection which can then be used to guide policy and inform the post-2015 development agenda, and here big data is acknowledged as a key component. At this meeting, four types of gender data gaps were identified: coverage across countries and/or regular country production, international standards to allow comparability, complexity, and granularity (sizeable and detailed datasets allowing disaggregation by demographic and other characteristics). Furthermore a series of big data types that have the potential to increase collection of gender data were identified:

- Mobile phone records: for example, mobile phone use and recharge patterns could be used as indicators of women's socioeconomic welfare or mobility patterns.
- Financial patterns: exploring engagement with financial systems.
- Online activity: for example, Google searches or Twitter activity which might be used to gain insights into women's maternal health, cultural attitudes, or political engagement.
- Sensing technologies: for example, satellite data which might be used to examine agricultural productivity, access to healthcare, and education services.
- Crowdsourcing: for example, disseminating apps to gain views about different elements of societies.

A primary objective of this meeting was to highlight that existing gender data gaps are large, and often reflect traditional societal norms, and that no data (or poor data) can have significant development consequences. Big data here has the

potential to transform the understanding of women's participation in work and communities. Crucially it was posited that while better data is needed to monitor the status of women in informal employment conditions, it is not necessarily important to focus on trying to extract more data but to make an impact with the data that is available to try and improve wider social, economic, and environmental conditions.

ILO, the UN, and Big Data

The aforementioned meeting represented one example of where the ILO has engaged with other stakeholders to not only acknowledge the importance of big data but begin to consider potential options for its use with respect to their agendas. However, as a UN agency, they partake in wider discussion with the UN regarding the importance of big data, as was seen in the 45th session of the UN Statistical Commission in March 2014 where the report of the secretary general on "big data and the modernization of statistical systems" was discussed (United Nations, 2014). This report is significant as it touches upon important issues, opportunities, and challenges that are relevant for the ILO with respect to the use of big data.

The report makes reference to the UN "Global Pulse" which is an initiative on big data established in 2009 which included a vision of a future where big data was utilized safely and responsibly. Its mission was to accelerate the adoption of big data innovation. Partnering with UN agencies such as the ILO, governments, academics, and the private sector, it sought to achieve a critical mass of implemented innovation and strengthen the adoption of big data as a tool to foster the transformation of societies.

There is a recognition that the national statistical system is essentially now subject to competition from other actors producing data outside of their system, and there is a need for data collection of national statistics to adjust in order to make use of the mountain of data now being produced almost continuously (and often automatically).

To make use of the big data, a shift may be required from the traditional survey-oriented collection of data to a more secondary data-focused orientation from data sources that are high in volume, velocity, and variety. Increasing demand from policy makers for real-time evidence in combination with declining response rates to national household and business survey means that organizations like the ILO will have to acknowledge the need to make this shift. There are a number of different sources of big data which may be potentially useful for the ILO: sources from administration, e.g., bank records; commercial and transaction data, e.g., credit card transactions; sensor data, e.g., satellite images or road sensors; tracking devices, e.g., mobile phone data; behavioral data, e.g., online searches; and opinion data, e.g., social media. Official statistics like those presented in ILO databases often rely on administrative data, and these are traditionally produced in a highly structured manner which can in turn limit their use. If administrative data was collected in real time, or in a more frequent basis, then it has the potential to become “big data.”

There are, however, a number of challenges related to the use of big data which face the UN, its agencies, and national statistical services alike:

- Legislative: in many countries, there will not be legislation in place to enable the access to, and use of, big data particularly from the private sector.
- Privacy: a dialogue will be required in order to gain public trust around the use of data.
- Financial: related to costs for access data.
- Management: policies and directives to ensure management and protection of data.
- Methodological: data quality, representativeness, and volatility are all issues which present potential barriers to the widespread use of big data.
- Technological: the nature of big data, particularly the volume in which it is often created meaning that some countries would need enhanced information technology.

An assessment of the use of big data for official statistics carried out by the UN indicates that there are good examples where it has been used, for

example, using transactional, tracking, and sensor data. However, in many cases, a key implication is that statistical systems and IT infrastructures need to be enhanced in order to be able to support the storage and processing of big data as it accumulates over time.

Modern society has witnessed an explosion of the quantity and diversity of real-time information known more commonly as big data, presenting a potential paradigm shift in the way official statistics are collected and analyzed. In the context of increased demand for statistics information, organizations recognize that big data has the potential to generate new statistical products in a timelier manner than traditional official statistical sources. The ILO, alongside a broader UN agenda to acknowledge the data revolution, recognizes the potential for future uses of big data at the global level, although there is a need for further investigation of the data sources, challenges and areas of use of big data, and its potential contribution to efforts working toward the “better work” agenda.

Cross-References

- [United Nations Educational, Scientific and Cultural Organization \(UNESCO\)](#)

Further Reading

- International Labour Organization. (2014a). *Key indicators of the labour market*. International Labour Organization. http://www.ilo.org/empelm/what/WCMS_114240/lang-en/index.htm. Accessed 10 Sep 2014.
- International Labour Organization. (2014b). *ILO databases*. International Labour Organization. <http://www.ilo.org/public/english/support/lib/resource/ilodatabases.htm>. Accessed 1 Oct 2014.
- International Labour Organization. (2014c). *ILOSTAT database*. International Labour Organization. http://www.ilo.org/ilostat/faces/home/statisticaldata?_afzLoop=342428603909745. Accessed 10 Sep 2014.
- United Nations. (2013). *Big data and modernization of statistical systems. Report of the Secretary-General*. United Nations. United Nations Economic and Social Council. Available at: <http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf>. Accessed 1 Dec 2014.
- United Nations. (2014). *UN global pulse*. United Nations. Available at: <http://www.unglobalpulse.org/>. Accessed 10 Sep 2014.

International Nongovernmental Organizations (INGOs)

Lázaro M. Bacallao-Pino
University of Zaragoza, Zaragoza, Spain
National Autonomous University of Mexico,
Mexico City, Mexico

In general terms, international nongovernmental organizations (INGOs) refer to private international organizations that are focused on solving various societal problems, often in developing countries. For example, INGOs might operate to provide access to basic services for the poor and to promote their interests, to provide relief to people who are suffering from disasters, or to work toward environmental protection and community development. INGOs have been included in what has been defined as the global civil society, sharing the same missions as other nongovernmental organizations (NGOs), but with an international scope. INGOs typically have outpost countries around the world, aimed at ameliorating a variety of problems.

INGOs have grown in number and have taken on increasingly important roles especially in the post-World War II era, to the extent that they have been considered central to and engines for issues such as the global expansion of human rights and the increasing environmental concerns and climate change and sustainable development. The importance of these global civil society actors has been recognized by a range of international actors and stakeholders. In fact, the United Nations (UN) has created mechanisms and rules for INGO participation in international conferences, arranging for consultations and clarifying their roles and functions as part of the international community. Across the board, INGOs are increasingly employing massive amounts of data in virtually all areas of concern to improve their work and decision-making.

Emergence and Main Characteristics of INGOs

The growth of INGOs has been explained based on various theoretical approaches. On the one hand, some perspectives offer “top-down” approaches, arguing that the rise of INGOs is associated with the degree of a country’s integration in world polity and international economy. On the other hand, “bottom-up” perspectives underline the evolution of democracy and the success of domestic economies as significant factors facilitating the growth of INGOs within certain countries. However, other approaches explain the rise of INGOs by taking into account a complex articulation of both economic and political factors and at two simultaneous levels of analysis: national and global.

The rising importance and presence of INGOs in the international policy arena over the second half of the twentieth century, and particularly since the 1990s, has been associated with factors such as the proliferation of complex humanitarian emergencies during the post-Cold War era, the divides produced by the withdrawal of state service provision as a result of the neoliberal privatization of public services, the failures of the schemes of government-to-government aid, the ineffectiveness and waste associated with the action of multilateral organizations, the growing distrust to politics and governments, and/or the emergence of evermore complex and challenging global problems. The convergence of these tendencies has created a space for the action of INGOs, with capacities and network structures in line with the emergent global-local scenario. As evidence of that importance, in February 1993, the UN Economic and Social Council (ECOSOC) set up an open-ended working group (OEWG) to review and update its agreements for consultation with NGOs and to establish consistent rules for the participation of these entities in international conferences organized by the UN.

Initially, INGOs were generally small and worked in particular places of the world, maintaining close relationships with target

beneficiaries. Later, they gained a positive reputation with donors and developing countries based on their actions. They also expanded, developing larger programs and activities, covering more technical and geographical areas. Significant expansion of INGOs took place in the 1980s, with funding peaking in contexts where donor conditions were relatively less rigorous. In such contexts, many INGOs put in practice processes of decentralization, increasing their networks by setting up regional or national offices in different countries.

INGOs are often defined in contrast to international governmental organizations (IGOs). INGOs have been described as any international organizations that are not established by inter-governmental agreements. Not constituted by states and not having structures of decision-making controlled by states are a defining characteristic of INGOs, although they may have contacts in governmental institutions or, as often happens, receive funding from states. Although sometimes described as apolitical in character and as separate from political parties, this does not mean that INGOs cannot take political stand. In fact, their actions can have important political implications at both international and domestic levels in a number of relevant issues – such as human rights – by, for instance, recommending certain policies and political actions. In this regard, they have been discussed in terms of “soft laws,” taking place and facilitating the participation of non-state actors such as INGOs in policy processes, influencing what has traditionally been framed as exclusive nation-state domains.

In this sense, INGOs have been included within the global “third sector.” The third sector refers to those entities that lie between the market and the state, separate from governmental structures and private enterprises. INGOs work outside both the global economy, a space dominated by transnational corporations and other financial institutions, and the global interstate system, configured by IGOs and typically centered around the UN. INGOs often are headquartered in developed

countries, while carrying out their activities in developing countries, operating across national borders and not identifying themselves as domestic actors.

However, INGOs are differentiated from other third sector actors, e.g., civil society organizations, based on aspects such as the INGO articulation in global consortia and their extensive global programmatic reach and the international arena in which they operate; their size and scope being much larger in terms of budgets, staffs, or operations; their greater organizational capacities and broader range of partnerships; and their higher profile, derived from the professionalism, credibility, and legitimacy that donors and the public associate with their actions. It is in this regard also that INGOs are making bigger commitments and investments in big data collection and use. As mentioned, INGO activities include a wide range of issues, from humanitarian and development assistance to human rights, gender, environment, poverty, education, research, advocacy, and international relief. To a great degree, INGOs are considered as spokespersons for global civil society on such themes since they become important spaces for collective action, as well as resources for participation in the global public sphere, contributing to the emergence and development of a global civic culture that includes related issues.

INGO Nonprofit Nature and Debates on Sources of Funding

A defining characteristic of INGOs is their not-for-profit nature. This is particularly important as it is related to the autonomy of INGOs. Sources of funding of INGOs can include individual donors who become members or partners of the organizations and philanthropy from private funds – such as private foundations – as well as donations from official development assistance programs provided by developed countries, churches and religious groups, artists, or some commercial activities, e.g., fair trade initiatives. Some claims

indicate that their nonprofit designations mean that INGOs may not conduct any operations that generate some private benefits. However, for example, INGOs are frequently professional organizations that have to, for instance, provide salaries to their employees and also sometimes participate in marketing campaigns to support their actions and agendas. Hence, their nonprofit nature only means that INGOs are differentiated from other private organizational actors, such as enterprises and corporations, because they are not explicitly for-profit organizations.

Big data analytics are being used to identify and track funding sources and donors. In that respect, INGO fundraising efforts have engaged big data to highlight aspects such as trends in individual giving that have decreased among younger generations or the preference of large private funding sources to seek similar large INGOs for partnership. Related analyses also underscore the consequences of dependency on government funds from official development assistance programs, which can be reduced as part of budgetary cuts during economic crises or political shifts. Besides dependence on fickle donor funds, whether public or private, possible shortages in organizational autonomy – conceptualized as the decision-making capacity of INGOs – have been noted due to funding source. That is, INGO operational and managerial autonomy may be constrained depending on funding source. Constraints associated with funding can include factors such as evaluation and performance controls, audit requirements, and various rules, regulations, and conditionalities. However, as a correlate of influence on INGOs exerted through funding and as an example of the abovementioned roles and impacts of INGOs' actions, they also can influence their funding sources through strategies such as exerting influence on the design and implementation of programs, contract negotiations, and revenue diversification or even by not applying for or accepting funds from certain sources that would constrain their autonomy.

From a more complex point of view, many debates on funding and INGO autonomy have

turned on a more general issue: the relationships between INGOs and governments. Of particular concern is rising bilateralization in the sense that an increasing amount of funding flows are directed toward specific countries and for particular purposes, while unrestricted funding has decreased. This tendency points not only to the matter of INGOs' autonomy but also to ideological debates on the interaction between governments and INGOs, summarized in two opposite positions: on the one hand, those who consider that the private nonprofit sector is the best mechanism for addressing social and economic needs, separating governments and INGOs, and, on the other hand, those who defend a strong welfare state, possibly minimizing the explicit need for nonprofit organizations depending on how they provide their services.

INGOs for Development Cooperation, Humanitarian Aid, and Human Rights

Three of the main areas of INGO action have been development cooperation, humanitarian aid, and human rights. Some of the largest INGOs are focused on one or more of these issues, and numerous approaches consider that the action of these organizations – and, in general, of nongovernmental agents – has been the engine for the global expansion and increasing importance of those topics, especially during the second half of the twentieth century.

INGOs focused on humanitarian aid have played a relevant role in large-scale humanitarian projects around the world, particularly during the last decades. They have provided emergency relief to millions of people and delivered important amounts of international humanitarian aid, assisting refugees, displaced persons, and persons living in conflict zones and scenarios of humanitarian crises or emergency, and long-term medical care to vulnerable populations. The actions of these INGOs are focused on aspects such as relief and rehabilitation, humanitarian mine action, and post-conflict recovery. Many of them also act in capacity building and cooperation between

authorities at different levels and implement activities in areas such as housing and small-scale infrastructure, income generation through grants and micro-finance, food security and agricultural rehabilitation and development, networking and capacity development, and advocating for equal access to healthcare worldwide. Among others, some important humanitarian INGOs are the Danish Refugee Council, CARE International, and Médecins Sans Frontières (MSF).

INGOs also have become increasingly relevant in the international development arena, with a rising amount of aid to developing countries, with budgets that, in the case of particularly large INGOs, have even surpassed those of some donor developed countries. Although INGOs involved in development cooperation assume diverse roles, there are significant similarities in their goals. Among their most frequent objectives are reducing poverty and inequality and the realization of rights, mainly for marginalized groups; the promotion of gender equality and social justice; the reinforcement of civil society and practices of democratic governance; and the protection of the environment. An increasing trend is to include research and learning processes as part of their strategies of action and as sources of data for establishing a more consolidated evidence base for both program experience and knowledge and policy influence. To mention only a few, some of the largest INGOs involved in development cooperation are BRAC, World Vision International, Oxfam International, and Acumen Fund.

Finally, the promotion and defense of human rights have been a particularly important action area for INGOs, becoming significant spaces for participation in the global human rights movement. Human rights INGOs, such as Amnesty International and Human Rights Watch, have been key agents in promoting human rights and in making known human rights violations. These INGOs oppose violations of rights such as freedom of religion or discrimination on the basis of sexual orientation, denouncing infringements related to gender discrimination, torture, military use of children, freedom of the press, political

corruption, and/or criminal justice abuses. By conducting campaigns on these themes, INGOs have drawn attention to human rights, mobilizing public opinion and pressuring governments to observe human rights in general.

Conclusion

In all of these situations, INGOs collect and utilize massive amounts of data for relevant planning, implementing, monitoring, and accountability activities. INGOs (and NGOs more generally) are putting data to effective use to measure and increase their impact, cut costs, identify and manage donors, and track progress.

Cross-References

- [Human Resources](#)
- [International Development](#)

Further Reading

- Boli, J., & Thomas, J. M. (1999). *Constructing world culture: International Nongovernmental Organizations since 1875*. Redwood City: Stanford University Press.
- Hobe, S. (1997). Global challenges to statehood: The increasingly important role of nongovernmental organizations. *Indiana Journal of Global Legal Studies*, 5(1), 191–209.
- McNeely, C. L. (1995). *Constructing the nation-state: International organization and prescriptive action*. Westport: Greenwood Press.
- Otto, D. (1996). Nongovernmental organizations in the United Nations system: The emerging role of international civil society. *Human Rights Quarterly*, 18(1), 107–141.
- Plakkot, V. (2015). 7 NGOs that are using data for impact and why you should use it too. <https://blog.socialcops.com/intelligence/best-practices/7-ngos-using-data-for-impact>.
- Powell, W. W., & Steinberg, R. (2006). *The nonprofit sector: A research handbook*. New Haven: Yale University Press.
- Tsutsui, K., & Min Wotipka, C. (2004). Global civil society and the international human rights movement: Citizen participation in human rights International Nongovernmental Organizations. *Social Forces*, 83(2), 587–620.

Internet Association, The

David Cristian Morar

Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Synonyms

[Internet Lobby](#); [Internet Trade Association](#); [Internet Trade Organization](#)

Introduction

The Internet Association is a trade organization that represents a significant number of the world's largest Internet companies, all of whom are based, founded, or ran in the United States of America. While issues such as net neutrality or copyright reform are at the forefront of their work, the Internet Association is also active in expressing the voice of the Internet industry in matters of Big Data. On this topic, it urges a commitment to status quo in privacy regulation and increased government R&D for innovative ways of enhancing the benefits of Big Data, while also calling for dispelling the belief that the web is the only sector that collects large data sets, as well as for a more thorough review of government surveillance. These proposals are underlined by the perspective that the government has a responsibility to protect the economic interests of the US industries, internationally, and a responsibility to protect the privacy of the American citizens, nationally.

Main Text

Launched in 2012 with 14 members and designed as the unified voice in Washington D.C. for the industry, the Internet Association now boasts 41 members and is dedicated, according to their statements, to protecting the future of the free and innovative Internet. Among these 41 members, some of the more notable include Amazon, AOL, Groupon, Google, Facebook, Twitter,

eBay, Yelp, IAC, Uber Technologies Inc, Expedia, and Netflix. As part of both their purpose and mission statements, the Internet Association believes that the decentralized architecture of the Internet, which it vows to protect, is what led it to become one of the world's most important engines for growth, economically and otherwise. The Association's representational role, also referred to as a lobbying, is portrayed as not simply an annex of Silicon Valley but as a voice of its community of users as well. The policy areas it promotes are explained with a heavy emphasis on the user and the benefits and rights the user gains.

The President and CEO, Michael Beckerman, a former congressional staffer, is the public face of the Internet Association, and he is usually the one that signs statements or comments on important issues on behalf of the members. Beyond their "business crawl" efforts promoting local businesses and their connection to, and success yielding from the Internet economy, the Association is active in many other areas. These areas include Internet freedom (nationally and worldwide) and patent reform, among others, with their most important concern being net neutrality. As Big Data is associated with the Internet, and the industry is interested in being an active stakeholder in related policy, the Association has taken several opportunities to make its opinions heard on the matter. These opinions can also be traced throughout the policies it seeks to propose in other connected areas.

Most notably, after the White House Office of Science and Technology Policy's (OSTP) 2014 request for information, as part of their 90-day review on the topic of Big Data, the Internet Association has released a set of comments that crystallize their views on the matter. Prior communications have also brought up certain aspects related to Big Data; however, the comments made to the OSTP have been the most comprehensive and detailed public statement to date by the industry on issues of Big Data, privacy, and government surveillance.

In matters of privacy regulation, the Association believes that the current framework is both robust and effective in relation to commercial entities. In their view, reform is mostly

necessary in the area of government surveillance, by adopting an update to the Electronic Communications Privacy Act (which would give service providers a legal basis in denying government requests for data that are not accompanied by a warrant), prohibiting bulk governmental collection of metadata from communications and clearly bounding surveillance efforts by law.

The Internet Association subscribes to the notion that the current regime for private sector privacy regulation is not only sufficient but also perfectly equipped to deal with potential concerns brought about by Big Data issues. The status quo is, in the acceptance of the Internet industry, a flexible and multilayered framework, designed for businesses that embrace privacy protective practices. The existing framework, beyond a sometimes overlapping federal-state duality of levels, also includes laws in place through the Federal Trade Committee that guard against unfair practices and that target and swiftly punish the bad actors that perpetrate the worst harms. This allows companies to harness the potential of Big Data within a privacy-aware context that does not allow or tolerate gross misconduct. In fact, the Association even cites the White House's 2012 laudatory comments on the existing privacy regimes, to strengthen its argument for regulatory status quo, beyond simply an industry's desire to be left to its own devices to innovate without major restrictions.

The proposed solutions by the industry would center on private governance mechanisms that include a variety of stakeholders in the decision-making process and are not, in fact, a product of the legislative system. Such actions have been taken before and, according to the views of the Association, are successful in the general sector of privacy, and they allow industry and other actors that are involved in the specific areas to have a seat at the table beyond the traditional lobbying route.

One part that needs further action, according to the views of the Association, is educating the public on the entire spectrum of activities that lead to the collection and analysis of large data sets. With websites as the focus of most privacy-related research, the industry advocates a more

consumer-oriented approach that would permeate the whole range of practices from understudied sectors to the Internet, centered around increasing user knowledge on how their data is being handled. This would allow the user to understand the entire processes that go on beyond the visible interfaces, without putting any more pressure on the industries to change their actions.

While the Internet Association considers that commercial privacy regulation should be left virtually intact, substantial government funding for research and development should be funneled into unlocking future and better societal benefits of Big Data. These funds, administered through the National Science Foundation and other instruments, would be directed toward a deeper understanding of the complexities of Big Data, including accountability mechanisms, de-identification, and public release. Prioritizing such government-funded research over new regulation, the industry believes that current societal benefits from commercial Big Data usage (ranging from genome research to better spam filters) would multiply in number and effect.

The Association deems that the innovation economy would suffer from any new regulatory approaches that are designed to restrict the free flow of data. In their view, not only would the companies not be able to continue with their commercial activities, which would hurt the sector, and the country, but the beneficial aspects of Big Data would suffer as well. Coupled with the revelations about the data collection projects of the National Security Agency, this would significantly impact the standing of the United States internationally, as important international agreements, such as the Transatlantic Trade and Investment Partnership with the EU, are in jeopardy, says the industry.

Conclusion

The Internet Association thus sees privacy as a significant concern with regard to Big Data. However, it strongly emphasizes governmental missteps in data surveillance, and offers an unequivocal condemnation of such actions,

while lauding and extolling the virtues of the regulatory framework in place to deal with the commercial aspect. The Association believes that current nongovernmental policies, such as agreements between users and service providers, or industry self-regulation, are also adequate, and promoting such a user-facing approach to a majority of privacy issues would continue to be useful. Governmental involvement is still desired by the industry, primarily through funding for what might be called basic research into the Big Data territory, as the benefits of this work would be spread around not just between the companies involved but also with the government, as best practices would necessarily involve governmental institutions as well.

Cross-References

- [De-identification/Re-identification](#)
- [Google](#)
- [National Security Agency \(NSA\)](#)
- [Netflix](#)

Further Reading

- The Internet Association. *Comments of the Internet Association in response to the White House Office of Science and Technology Policy's Government 'Big Data' Request for Information*. http://internetassociation.org/wp-content/uploads/2014/03/3_31_-2014_The-Internet-Association-Comments-Regarding-White-House-OSTP-Request-for-Information-on-Big-Data.pdf. Accessed July 2016.
- The Internet Association. *Comments on 'Big Data' to the Department of Commerce*. <http://internetassociation.org/080614comments/>. Accessed July 2016.
- The Internet Association. *Policies*. <https://internetassociation.org/policy-platform/protecting-internet-freedom/>. Accessed July 2016.
- The Internet Association. *Privacy*. <http://internetassociation.org/policies/privacy/>. Accessed July 2016.
- The Internet Association. *Statement on the White House Big Data Report*. <http://internetassociation.org/050114bigdata/>. Accessed July 2016.
- The Internet Association. *The Internet Association's Press Kit*. <http://internetassociation.org/the-internet-associations-press-kit/>. Accessed July 2016.
- The Internet Association. *The Internet Association Statement on White House Big Data Filed Comments*. <http://internetassociation.org/bigdatafilingstatement/>. Accessed July 2016.

Internet Lobby

- [Internet Association, The](#)

Internet of Things (IoT)

Erik W. Kuiler

George Mason University, Arlington, VA, USA

The Internet of Things (IoT) is a global computing-based network infrastructure, comprising uniquely identifiable objects embedded in entities connected via the Internet that can collect, share, and send data and act on the data that they have received. IoT defines how these objects will be connected through the Internet and how they will communicate with other objects by publishing their capabilities and functionalities as services and how they may be used, merging the digital (virtual) universe and the physical universe.

The availability of inexpensive computer chips; advances in wireless sensor networks technologies; the manufacture of inexpensive radio-frequency identification (RFID) tags, sensors, and actuators; and the ubiquity of wireless networks have made it possible to turn anything, such as telephony devices, household appliances, and transportation systems, into IoT participants. From a tropological perspective, IoT represents physical objects (*things*) as virtual entities that inhabit the Internet, thereby providing a foundation for cloud-based big data analytics and management.

Conceptually, the IoT comprises a framework with several interdependent tiers:

Code tier – the code tier provides the foundation for IoT, in which each object is assigned a unique identifier to distinguish it from other IoT objects.

Identification and recognition tier – the identification and recognition tier comprises, for example, RFID tags, IR sensors, or other sensor networks. Devices in this tier gather information about objects from the sensor devices

linked with them and convert the information into digital signals which are then passed onto the network tier for further action.

Network tier – the devices in the network tier receive and transmit the digital signals from devices in the identification and recognition tier and transmit it to the processing systems in the middleware tier through various media, e.g., Bluetooth, WiMaX, Zigbee, GSM, 3G, etc., using the appropriate protocols (IPv4, IPv6, MQTT, DDS, etc.).

Middleware tier – devices in this tier process the information received from the sensor devices. The middleware tier includes the cloud-based ubiquitous computing functions that ensure direct access to the appropriate data stores for processing.

Application tier – software applications in this tier instantiate support for IoT-dependent applications, such as smart homes, smart transportation systems, smart and connected cities, etc.

Business tier – software applications in this tier support IoT-related research and development as well as the evolution of business strategies, models, and products.

Relative to these various tiers, the IoT is typically discussed in terms of technological advances and improvements to the human condition. However, there are issues that require more critical review and consideration. For example, security is a principal concern. IoT security breaches may take the form of unauthorized access to RFID, breaches of sensor-nodes security, cloud-based computing abuse, etc. Also, to ensure reliability and efficacy, IoT devices and networks must ensure interoperability – technical interoperability, syntactical interoperability, semantic interoperability, and organizational interoperability – but, again, that raises further security issues. In the ubiquitous IoT environment, there are no clear ways to establish and secure human anonymity. In addition to deliberate (positive or negative) purposes, the inadvertent dissemination of personally identifiable information (PII), privacy information, and similar information occurs all too frequently, and oversight of related devices

and information is increasingly difficult. In fact, miniaturization also plays a role in this regard. Many IoT personal devices are reduced to the point of invisibility, minimizing transparency to human overview and management.

The IoT comprises the network of devices embedded in everyday objects that are enabled to receive, act on, and send data to each other via the Internet. For efficacy and efficiency, the IoT relies on a multi-tiered framework that ensues syntactic conformance, semantic congruence, and technological reliability. In general terms, it can be framed in terms of autonomous agency relative to the increasing prevalence, reliance, and risks of (unintended spontaneous) intervention in human events. As such, the IoT also reflects ontological ambiguity, blurring distinctions between human beings, natural objects, and artifacts as parts of the broader smart and digitized environment.

Cross-References

► [Data Streaming](#)

Further Reading

- Bandyopadhyay, D., & Sen, J. (1995). Internet of Things – Applications and challenges in technology and standardization. *Wireless Personal Communications*, 58 (1), 49–69.
- Cheng, X., Zhang, M., & Sun, F. (2012). Architecture of internet of things and its key technology integration based on RFID. In *IEEE fifth international symposium on computational intelligence and design* (pp. 294–297).
- European Research Cluster on the Internet of Things (IERC). (2015). *Internet of Things IoT semantic interoperability: research challenges, best practices, recommendations and next steps*. Retrieved from: <http://www.internet-of-things-research.eu/>.
- Voas, J. (2016). NIST special publication 800-183: Networks of ‘Things.’ Retrieved from: Networks of ‘Things’ (nist.gov).
- Wu, M., Lu, T.-L., Ling, F.-Y., Sun, L., & Du, H.-Y. (2010). Research on the architecture of Internet of things. In *Advanced computer theory and engineering* (pp. 484–487).
- Zhang, Y. (2011). Technology framework of the Internet of Things, and its application. In *IEEE third international conference on electronics and communication engineering* (pp. 4109–4112).

Internet Trade Association

► [Internet Association, The](#)

Internet Trade Organization

► [Internet Association, The](#)

Internet: Language

Marcienne Martin

Laboratoire ORACLE [Observatoire Réunionnais des Arts, des Civilisations et des Littératures dans leur Environnement] Université de la Réunion Saint-Denis France, Montpellier, France

It is the Arpanet network that will be at the origin the Internet. It was established in 1969 by the United States Department of Defense. As Mark (1999) mentions, the Internet has several characteristics including the decentralization of transmissions, which means that when a line of communication becomes inoperable the two remote machines will search for a new path to transfer the data (the circuit can start on the East Coast of Canada, through the province of Ontario, and finally lead to Saskatchewan). Arpanet has a special mode of communication between computers; Internet Protocol [IP]. [IP] works as a sort of electronic envelope into which data are put. In January 1994, the vice president of the United States, Al Gore, for the first time used the term “information highway” to describe the American project to construct a national network of modern communication. The network as we know it has been promoted by a group of research institutes and universities under the direction of Professor Clever. It consisted of five interconnected supercomputers located in different geographic areas. According to Mark (1999), “Computers are able to collaborate, forming interconnected cells of

an electronic brain.” Dr. Clever has proposed combining Arpanet, NSF, Bitnet, Usenet, and all other networks into a single entity called the Internet. The Internet has become “a homogeneous material resulting from a large number of individual networks that are composed of many heterogeneous computer systems (individuals, businesses, government institutions)” (Mark 1999).

Tanenbaum (2001) shows that the reticular structure of digital society is structured around six levels of language, that is, among others, the machine language, the programming language, the language used by the user of this media, or natural language, let alone the new language illustrated by the smiley. Tanenbaum specifies that every language is built on its predecessor so that we can see a computer as a multilayer stack or levels. The language of the bottom is the simplest, the top one the most complex. Machine language, which is the structural basis of the Internet, is a binary sequence (e.g., 0111 1001) which can only be understood by experts and, therefore, is unusable as such in everyday communication. It is something of a raw material to pass through a number of transformations in order to be used.

A great number of researchers agree on the fact that this new digital paradigm, which is part of the Internet, forms the basis for a transformation in social behavior that affects a large proportion of the world population. The analysis of the digital society is different from one researcher to another. Marshall McLuhan (2001) mentions the Internet as a global village, without borders, without law, and without constraint. For Wolton (2000) the screen of the computer will simplify the communication between human beings and make it more direct and transparent, while the computer system will be more regulated and more closed and more coded. Wolton mentions that in civil society there is never a transparent social relation. Furthermore, the author specifies that access to knowledge and information is the source of the revival of inequality. The risk is that there is a place for everyone, but yet every one remains in their place. Proulx (2004) found that communication in the digital society transforms the space-time relation. The user has access to information anytime and

anywhere, which results in a generalization of consultation of sites located in different parts of the world in delayed time. Contrary to users of traditional media (for example, television and radio broadcasting), the Internet user is an innovator in the management of a written code that uses the style and syntax of an oral code. This innovative character takes into account what is already there. So communication through this media is at the origin of a new language using the alphanumeric signs and symbols located on the keys of the physical or digital keyboard and this, in the context of an innovative semantic context. While the user is in front of their screen, they do not see their interlocutors. This staging of reality refers only to the imagination of the Net surfer and their interpretation of the situation communication. Unlike the television in which the subject is rather passive – it can use the “zapping” or turn off the TV – the Internet user can break off a conversation if they find it inappropriate, without giving any justification, which is not the case in an exchange of traditional communication. The rules of etiquette (good manners), even if they are advocated on the Web, may, however, be ignored. In civil society, a speaker who would not make dialogic openings and closures inherent to their culture would be sanctioned by the rejection whatsoever from the caller and/or their group of belonging.

Communication in humans via conversational exchanges has been the subject of numerous studies. A specific mode of verbal interaction, that is conversation, was studied in particular by Kerbrat-Orecchioni (1996); she shows that the principal characteristics are the implication of a limited number of participants playing roles not predetermined, benefiting normally of the same rights and duties and having the pleasure of conversing; conversation has a familiar and improvised nature whatsoever at the themes, the duration of the exchange, the order of the speeches. As specified by the author: “The interaction is symmetric and egalitarian.” However, some parameters are involved in the proper conduct of this type of interaction: it is the sharing to the same linguistic and cultural heritage. Indeed, the reports in the world can differ from one group

to another and the semantic content of a particular meaning can take different values, generating situations of misunderstanding or even of conflict.

However, when referring to the space of the Internet, one has to consider a new order of the number of participants involved in the conversational exchange. Indeed, the digital technology that forms the basis of this medium permits an unlimited number of Internet users to connect to a particular chat room. Some chat rooms can display a large number of participants. This means that we are far from being faced with conversational patterns in real life and for which such exchanges would be doomed to fail. In the digital society, each user is alone behind their machine, and it is all of these units that form an informal group composed by the participants of a particular chat room. Furthermore, the perception that Net surfers can have concerning the number of speakers involved in the activity in which they participate may be misleading. That is why to overcome the problem posed by the large number of users connected at the same time, in the same chat room; discussions were set up called “private” and expressed by the abbreviation “PV.” Regarding the exchange turns, here we are in the case where the digital structure that underlies the Internet medium supports the management of this event. Thus, the electrical impulses that work to create equations, so-called Boolean, operate in consecutive ranking. This order is reflected in the upper layers of more sophisticated programming languages than at the level of Net surfers. Turn-taking of speakers is, therefore, not managed by users but by the computer.

Moreover, the only organs solicited within the framework of communicative exchanges on the Internet are eyes for reading on the screen and writing the message on the keyboard, as well as the touch when using the keyboard keys; this implies that in this particular universe there is the absence of any kinesics manifestation, that the opening of a dialogue takes place on the basis of a soliloquy, that the establishment of a single proxemic distance is common to all Internet users, namely, the physical distance that separates them from the computer tool.

The New Language on Internet

The field of writing seems to correspond to a widening of the field of speech both on a spatial and temporal level. Boulanger (2003) contends that through the medium of limited sounds and possible actions, man has forged a speech organized and filled with meaning. For Leroi-Gourhan, anthropologist, the history of writing begins with tracings and visuals of the end of the Mousterian period, around 50,000 BC, and then it propagates around 30,000 BC. These tracings open to interpretation would have served as a mnemonic support. This proto writing consisted of incisions (lines, points, grooves, sticks, etc.) regularly spaced and formed in stones or bones. This is the development of external oral code through the writing support.

Referring to the language used on Internet means evoking a hybrid structure, first take into account written support to express a message and the other, makes extensive use of terms used in spoken in the lexical-semantic phrases. Thus, the identification and analysis of discursive sequences show that the form of rebus with the use of logograms has been adopted, such as numbers and the sign arrobas: @; these characters are at the origin of the phonetic support of written medium objects and of its oral version, rapid writing which is the abbreviation of words like *Pls* (please), which reduce the message to its phonetic transcription as *ID* (idea), or use a mixture of rapid writing and phonetic transcription. The personal creation governs linguistic innovation in the Internet; it is manifested in rebus, rapid writing, and phonetic reduction, etc. So the puzzle is made, often logograms, phonemes, stylistic figures, etc. as *C**** (cool). Poets have thus used as Queneau (1947) with the use of these stylistic figures in writing poems.

In addition, the writing on the Internet uses semantic keys to the image of Chinese characters, on one hand would serve to create complex logograms and on the other hand, would initialize particular field semantics (Martin 2010). These keys are at the origin of basic pictographs composed of a simple graph; in their combined form, these graphs result in more complex pictograms.

Some of these graphs are shown in Table 1; for each of them, the basic icons are those listed in the table of characters on the keyboard used. Thus, the semantic field of facial expressions has several keys that initiate eyes, mouth, and nose, respectively, as we can see in Table 1. Simplified pictographs are unambiguous and monosemic, but in their more complex version the reading of these icons request the use of a legend. Usually their creators add a small explanatory text. Moreover, these symbols punctuate the linguistic discourse, due to the inability to compensate paraverbal and nonverbal exchange set up during the usual conversations implemented in civil society.

Identity on the Internet

In the image of the complexity of the universe including humans and the organizations in which they are part, subsuming various paradigms, such the divine, the human, and the objectal, etc., nomination is a fascinating phenomenon but difficult, almost impossible, to define in its entirety. Number of parameters and factors modify both fixed and variable components. By the consciousness of being in the world, while questioning the strangeness live to die, humans take place in reality by naming them. Anthroponomy takes part in this process; its organization reflects the culture of which it is part. Patronymic and first names have in common the quality of *nomen verum* (veritable name) (Laugaa 1986). Ghasarian (1996) emphasizes with the patronymic as the noun of relationship that an individual receives at birth, demonstrating its identity. Moreover, the first name would be similar to that pseudonym its actualization occurs in the synchronic time and

Internet: Language, Table 1 Semantic field of expressions of the face

Eyes			
:	;	‘	,
Open eyes	Nod	Left eyebrow	Right eyebrow
Expressive gestures of the mouth			
)	(	!	<
Smile	Pout	Indifference	Disappointment

not in the diachronic time (transgenerational) as to the surname.

As opposed to the surname, pseudonyms do not infer any genealogical connection. However, if the construction is personal creation order, it remains highly contextualized. Thus autonyms created by users for the need to surf the Internet space while preserving the confidentiality of their privacy will be motivated by the particularity of this media. However, the fact to evoke the place of the individuals in the genealogical chain implicitly refers to the construction of their identity, of their groups of belonging and/or of opposition, and finally to the definition of their status. However, the construction of a pseudonym on the Internet actualizes new social habits that will depend on both the personal choice of the user and of the virtual society they wish to join.

Digital media, including network structures (networking), form the basis of the function of *nomen falsum* (false name) is plural. A nickname is rigid and an identity marker at a given time. Indeed, because of the configuration of the computer system which is running according to a binary mode, homonymic names are only recognized as a single occurrence. However, users of the Internet can change their pseudonym ad libitum. The nominal sustainability is not correlated to the holder of such surnames as is the case in civil society where the law lays down strict rules for the official nomination. The creation of the pseudonym is made from data taken, among others, in the private life of the Net surfer (Martin 2006, 2012).

In civil society, anthroponomy sets the social being within a group. In order to join discussion forums or chat rooms, the internet user has to choose a pseudonym. However, using a pseudonym on the Web is not done for the sole purpose of naming an individual. The main feature of the pseudonym on the Internet is its richness in terms of creativity. Moreover, some nicknames become discursive spaces where users claim positions already taken, issue opinions, or express their emotions. A *nomen falsum* can serve the user's speech. Both designator and discursive unity, the pseudonym amplifies by synthesizing the speech of the user. It can also act as an

emotional vector. Marker of identity at the base, it is an anthroponym called pseudonym. Like the mask, it has a plural vocation.

Social networks are an extension of chat rooms with more personalized communication modalities. One example is the *Facebook* social network that allows any user of the Web to create a personal space with the ability to upload photos and videos, to write messages on an interface (the wall) which can be consulted by relatives and friends, or by all the members of the social network to the extent that the user accepts this possibility. It was in 2004 that the social network *Facebook* was created by Zuckerberg and his fellow students at Harvard University, Eduardo Saverin, Dustin Moskovitz, and Chris Hughes. Other social networks like *LinkedIn*, established in 2003, belong to professional online social networks; their network structure works from several levels of connection: direct contacts, contacts of direct contacts, and then the contacts to the second degree. There are also social networks like *Twitter* whose characteristic is sending messages limited to 140 characters. *Twitter* took its name from the company Twitter Inc. creator of this social network; it is a blogging platform that allows a user to send free short messages called tweets on the internet, instant messaging or SMS.

Social networks are a way to enrich the lives of Internet users by virtually meeting with users sharing the same tastes, the same opinions, etc. Social networks can be at the origin of the reorientation of political or social opinion. Nevertheless, connecting to *Facebook* can often be the cause of a form of addiction, since acting on this social network allows users to create a large network of virtual friends, which can also affect the user's image. Thus, having a lot of friends may refer to an overvalued self-image, while the opposite may result in a devaluation of one's image.

The study of the territory of the internet and social practices that it induces, refers, on the one hand, to the Internet physical territory occupied by the user, that is to say, a relationship between the keyboard and the screen on a space belonging to what Hall defines as "intimate distance" and, on the other hand, the symbolic territory that

registered the other in a familiar space. These different modes of running of the pseudonym are correlated to the development of *nomen falsum* (nickname) on the personal territory of the Net surfer, both physical and symbolic, and more specifically in the context of its intimate sphere, which has profound implications concerning the relations between the communication of Internet users. The Internet is a breeding ground where creativity takes shape and grows exponentially. These are the new locations of speech in which the exchange engaged by users can have repercussions in civil society.

Further Reading

- Boulanger, J.-C. (2003). *Les inventeurs de dictionnaires*. Ottawa: Les presses de l'Université d'Ottawa.
- Ghasarian, C. (1996). *Introduction à l'étude de la parenté*. Paris: Editions du Seuil.
- Hall, T. E. (1971). *La dimension cachée, édition originale*. Paris: Seuil.
- Kerbrat-Orecchioni, C. (1996). *La conversation*. Paris: Seuil.
- Laugaa, M. (1986). *La pensée du pseudonyme*. Paris: PUF.
- Leroi-Gourhan, A. (1964). *Le Geste et la Parole, Technique et langage*. Paris: Albin Michel.
- Mark, T. R. (1999). *Internet, surfez en toute simplicité sur le plus grand réseau du monde*. Paris: Micro Application.
- Martin, M. (2006). *Le pseudonyme sur Internet, une nomination située au carrefour de l'anonymat et de la sphère privée*. Paris: L'Harmattan.
- Martin, M. (2010). *Dictionnaire des pictogrammes numériques et du lexique en usage sur Internet et sur les téléphones portables*. Paris: L'Harmattan.
- Martin, M. (2012). *Se nommer pour exister – L'exemple du pseudonyme sur Internet*. Paris: L'Harmattan.
- McLuhan, M., & Fiore, Q. (2001). *The medium is the MESSAGE*. Hamburg/Berkeley: Gingko Press.
- Proulx, S. (2004). *La révolution Internet en question*. Montréal: Québec Amérique.
- Queneau, R. (1947). *Exercices de style*. Paris: Gallimard.
- Tanenbaum, A. (2001). *Architecture de l'ordinateur*. Paris: Dunod.
- Wolton, D. (2000). *Internet et après?* Paris: Flammarion.

Italy

Chiara Valentini

Department of Management, Aarhus University,
School of Business and Social Sciences, Aarhus,
Denmark

Introduction

Italy is a Parliamentary republic in southern Europe. It has a population of about 60 million people of which, 86.7%, are Internet users (Internet World Stat 2017). Public perception of handling big data is generally very liberal, and the phenomenon has been associated with more transparency and digitalized economic and social systems. The collection and processing of personal data have been increasingly used to counter tax evasion which is one of the major problems of Italian economy. The Italian Revenue Agency is using data collected through different private and public data collectors to cross-check tax declarations (DPA 2014a).

According to the results of a study on Italian companies' perception of big data conducted by researchers at the *Big Data Analytics & Business Intelligence Observatory* of Milan Polytechnic, more and more companies (+22% in 2013) are interested in investing in technologies that allow to handle and use big data. Furthermore, the number of companies seeking professional managers that are capable of interpreting data and assisting senior management on decision-making is also increasing. Most of the Italian companies (76% of 184 interviewed) claim that they use basic analytics strategically and another 36% use more sophisticated tools for forecasting activities (Mosca 2014, January 7).

Data Protection Agency and Privacy Issues

Despite the positive attitude and increased use of big data by Italian organizations, an increasing

Invisible Web, Hidden Web

► [Surface Web vs Deep Web vs Dark Web](#)

public expectation for privacy protection has emerged as a result of raising debates on personal data, data security, and protection in the whole European Union. In the past years, the Italian Data Protection Authority (DPA) reported several instances of data collection of telephone and Internet communications of Italian users which may have harmed Italians' fundamental rights (DPA 2014b). Personal data laws have been developed as these are considered important instruments for the overall protection of fundamental human rights, thereby adding new legal specifications to the existing privacy framework. The first specific law on personal data was adopted by the Italian Parliament in 1996 and this incorporated a number of guidelines already included in the European Union 1995 *Data Protection Directive*. At the same time, an independent authority, the Italian Data Protection Authority (*Garante per la protezione dei dati personali*), was created in 1997 to protect fundamental rights and freedoms of people when personal data are processed. The Italian Data Protection Authority (DPA) is run by a four-member committee elected by the Italian Parliament for a seven-year mandate (DPA 2014a).

The main activities of DPA consist of monitoring and assuring that organizations comply with the latest regulations on data protection and individual privacy. In order to do so, DPA carries out inspections on organizations' databases and data storage systems to guarantee that their requirements for preserving individual freedom and privacy are of high standards. It checks that the activities of the police and the Italian Intelligence Service comply with the legislation, reports privacy infringements to judicial authorities, and encourages organizations to adopt codes of conduct promoting fundamental human rights and freedom. The authority also handles citizens' reports and complaints of privacy loss or any misuse or abuse of personal data. It bans or blocks activities that can cause serious harm to individual privacy and freedom. It grants authorizations to organizations and institutions to have access and use sensitive and/or judicial data. Sensitive and judicial data concern, for instance, information on a person's

criminal records, ethnicity, religion or other beliefs, political opinions, membership of parties, trade unions and/or associations, health, or sex life. Access to sensitive and judicial data is granted only for specific purposes, for example, in situations where it is necessary to know more about a certain individual for national security reasons (DPA 2014b).

The DPA participates to data protection activities involving the European Union and other international supervisory authorities and follows existing international conventions (Schengen, Europol, and Customs Information System) when regulating Italian data protection and security matters. It carries out an important role in increasing public awareness of privacy legislation and in soliciting the Italian Parliament to develop legislation on new economic and social issues (DPA 2014b). The DPA has also formulated specific guidelines on cloud computing for helping Italian businesses. Yet, according to this authority, these cloud computing guidelines require that Italian laws are updated to be fully effective in regulating this area. Critics indicate that there are limits in existing Italian laws concerning the allocation of liabilities, data security, jurisdiction, and notification of infractions to the supervisory authority (Russo 2012).

Another area of great interest for the DPA is the collection of personal data via video surveillance both in the public and in the private sector. The DPA has acted on specific cases of video surveillance, sometimes banning and other times allowing it (DPA 2014c). For instance, the DPA reported to have banned the use of webcams in a nursery school to protect children's privacy and to safeguard freedom of teaching. It banned police headquarters to process images collected via CCTV cameras installed in streets for public safety purposes because such cameras also captured images of people's homes. The use of customers' pre-recorded, operator-unassisted phone calls for debt collection purposes is among those activities that have been prohibited by this authority. Yet, the DPA permits the use of video surveillance in municipalities for counter-vandalism purposes (DPA 2014b).

Conclusion

Overall, Italy is advancing with the regulation of big data phenomenon following also the impetus given by the EU institutions and international debates on data protection, security, and privacy. Nonetheless, Italy is still lagging behind many western and European countries regarding the adoption and development of frameworks for a full digital economy. According to the *Networked Readiness Index 2015* published by the World Economic Forum, Italy is ranked 55th. As indicated by the report, Italy's major weakness is still a political and regulatory environment that does not facilitate the development of a digital economy and its innovation system (Bilbao-Osorio et al. 2014).

Cross-References

- [Cell Phone Data](#)
- [Data Security](#)
- [European Union](#)
- [Privacy](#)

References

- Bilbao-Osorio, B., Dutta, S. & Lanvin, B. (2014). The global information technology report 2014. Reword and risks of big data. *World Economic Forum*. http://www3.weforum.org/docs/WEF_GlobalInformationTechnology_Report_2014.pdf. Accessed 31 Oct 2014.
- DPA (2014a). *Summary of key activities by the Italian DPA in 2013*. <http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/3205017>. Accessed 31 Oct 2014.
- DPA (2014b). *Who we are*. http://www.garanteprivacy.it/web/guest/home_en/who_we_are. Accessed 31 Oct 2014.
- DPA. (2014c) "*Compiti del Garante*" [*Tasks of DPA*]. <http://www.garanteprivacy.it/web/guest/home/autorita/compiti>. Accessed 31 Oct 2014.
- Internet World Stat (2017). Italy. <http://www.internetworldstats.com/europa.htm>. Accessed 15 May 2017.
- Mosca, G. (2014, January 7). Big data, una grossa opportunità per il business, se solo si sapesse come usarli. La situazione in Italia. La Stampa. <http://www.ilsole24ore.com/art/tecnologie/2014-01-07/big-data-grossa-opportunita-il-business-se-solo-si-sapesse-come-usarli-situazione-italia-110103.shtml?uuid=ABuGM6n>. Accessed 31 Oct 2014.
- Russo, M. (2012). Italian data protection authority releases guidelines on cloud computing. In McDermott Will & Emery (Eds.), *International News* (Focus on Data Privacy and Security, 4). <http://documents.lexology.com/475569eb-7e6b-4aec-82df-f128e8c67abf.pdf>. Accessed 31 Oct 2014.

Journalism

Brian E. Weeks¹, Trevor Diehl², Brigitte Huber²
and Homero Gil de Zúñiga²

¹Communication Studies Department, University
of Michigan, Ann Arbor, MI, USA

²Media Innovation Lab (MiLab), Department of
Communication, University of Vienna, Wien,
Austria

The Pew Research Center notes that journalism is a mode of communication that provides the public verified facts and information in a meaningful context so that citizens can make informed judgments about society. As aggregated, large-scale data have become readily available and the practice of journalism has increasingly turned to big data to help fulfill this mission. Journalists have begun to apply a variety of computational and statistical techniques to organize, analyze, and interpret these data, which are then used in conjunction with traditional news narratives and reporting techniques. Big data are being applied to all facets of news including politics, health, the economy, weather, and sports.

The growth of “data-driven journalism” has changed many journalists’ news gathering routines by altering the way news organizations interact with their audience, providing new forms of content for the public and incorporating new methodologies to achieve the objectives of journalism. Although big data offer many

opportunities for journalists to report the news in novel and interesting ways, critics have noted data journalism also faces potential obstacles that must be considered.

Origins of Journalism and Big Data

Contemporary data journalism is rooted in the work of reporters like Philip Meyer, Elliot Jaspín, Bill Dedman, and Stephen Doig. In his 1973 book, Meyer introduced the concept of “precision journalism” and advocated applying social science methodology to investigative reporting practices. Meyer argued that journalists needed to employ the same tools as scientific researchers: databases, spreadsheets, surveys, and computer analysis techniques.

Based on the work of Meyer, computer-assisted reporting developed as a niche form of investigative reporting by the late 1980s, as computers became smaller and more affordable. A notable example from this period was Bill Dedman’s Pulitzer Prize winning series “The Color of Money.” Dedman obtained lending statistics on computer tape through the federal Freedom of Information Act. His research team combined that data with demographic information from the US Census. Dedman found widespread racial discrimination in mortgage lending practices throughout the Atlanta metropolitan area.

Over the last decade, the ubiquity of large, often free, data sets has created new opportunities

for journalists to make sense of the world of big data. Where precision journalism was once the domain of a few investigative reporters, data-driven reporting techniques are now a common, if not necessary, component of contemporary news work. News organizations like *The Guardian*, *The New York Times*' *Upshot*, and *The Texas Tribune* represent the mainstream embrace of big data. Some websites, like Nate Silver's *FiveThirtyEight*, are entirely devoted to data journalism.

How Do Journalists Use Big Data?

Big data provide journalists with new and alternative ways to approach the news. In traditional journalism, reporters collect and organize information for the public, often relying on interviews and in-depth research to report their stories. Big data allow journalists to move beyond these standard methods and report the news by gathering and making sense of aggregated data sets. This shift in methods has required some journalists and news organizations to change their information-gathering routines. Rather than identifying potential sources or key resources, journalists using big data must first locate relevant data sets, organize the data in a way that allows them to tell a coherent story, analyze the data for important patterns and relationships, and, finally, report the news in a comprehensible manner. Because of the complexity of the data, news organizations and journalists are increasingly working alongside computer programmers, statisticians, and graphic designers to help tell their stories.

One important aspect of big data is visualization. Instead of writing a traditional story with text, quotations, and the inverted-pyramid format, big data allow journalists to tell their stories using graphs, charts, maps, and interactive features. These visuals enable journalists to present insights from complicated data sets in a format that is easy for the audience to understand. These visuals can also accompany and buttress news articles that rely on traditional reporting methods.

Nate Silver writes that big data analyses provide several advantages over traditional

journalism. They allow journalists to further explain a story or phenomenon through statistical tests that explore relationships, to more broadly generalize information by looking at aggregate patterns over time and to predict future events based on prior occurrences. For example, using an algorithm based on historical polling data, Silver's website, *FiveThirtyEight* (formerly hosted by the *New York Times*), correctly predicted the outcome of the 2012 US presidential election in all 50 states. Whereas methods of traditional journalism often lend themselves to more microlevel reporting, more macrolevel and general insights can be gleaned from big data.

An additional advantage of big data is that, in some cases, they reduce the necessary resources needed to report the story. Stories that would otherwise have taken years to produce can be assembled relatively quickly. For example, WikiLeaks provided news organizations nearly 400,000 unreleased US military reports related to the war in Iraq. Sifting through these documents using traditional reporting methods would take a considerable amount of time, but news outlets like *The Guardian* in the UK applied computational techniques to quickly identify and report the important stories and themes stemming from the leak, including a map noting the location of every death in the war.

Big data also allow journalists to interact with their audience to report the news. In a process called crowdsourcing the news, large groups of people contribute relevant information about a topic, which in the aggregate can be used to make generalizations and identify patterns and relationships. For example, in 2013 the *New York Times* website released an interactive quiz on American dialects that used responses to questions about accents and phrases to demonstrate regional patterns of speech in the US. The quiz became the most visited content on the website that year.

Data Sets and Methodologies

Journalists have a multitude of large data sets and methodologies at their disposal to create news

stories. Much of the data used is public and originates from government agencies. For example, the US government has created a website, data.gov, which offers over 100,000 datasets in a variety of areas including education, finance, health, jobs, and public safety. Other data, like the WikiLeaks reports, were not intended to be public but became primary sources of big data for journalists. News organizations can also utilize publicly available data from private Internet companies like Google or social networking sites such as Facebook and Twitter to help report the news.

Once the data are secured, journalists can apply numerous techniques to make sense of the data. For example, at a basic level, journalists could get a sense of public interest about a topic or issue by examining the volume of online searches about the topic or the number of times it was referenced in social media. Mapping or charting occurrences of events across regions or countries also offers basic descriptive visualizations of the data. Journalists can also apply content or sentiment analyses to get a sense of the patterns of phrases or tone within a set of documents. Further, network analyses could be utilized to assess connections between points in the data set, which could provide insights on the flow or movement of information, or on power structures.

These methods can be combined to produce a more holistic account of events. For example, journalists at the Associated Press used textual and network analysis to examine almost 400,000 WikiLeaks documents related to the Iraq war that identified related clusters of words used in the reports. In doing so, they were able to demonstrate patterns of content within the documents, which shed previously unseen light on what was happening on the ground during the war.

Computer algorithms, and self-taught machine learning techniques, also play an important role in the big data journalistic process. Algorithms can be designed to automatically write news stories, without a human author. These automated “robot journalists” have been used to produce stories for news outlets like the *Associated Press* and *The Los Angeles Times*. Algorithms have also changed the way news is delivered, as news aggregators

like Google News employ these methods to collect and provide users personalized news feed.

Limitations of Big Data for Journalism

Although big data offer numerous opportunities to journalists reporting the news, scholars and practitioners have both highlighted several potential general limitations of these data. As much as big data can help journalists in their reporting, they need to make an active effort to contextualize the information. Big data storytelling also elicits moral and ethical concerns with respect the data collection of individuals as aggregated information. These reporting techniques also need to bear in mind potential data privacy transgressions.

Cross-References

- [Computational Social Sciences](#)
- [Data Visualization](#)
- [Digital Storytelling, Big Data Storytelling](#)
- [Information Society](#)
- [Interactive Data Visualization](#)
- [Open Data](#)

Further Reading

- Pew Research Center. *The core principles of journalism*. <http://www.people-press.org/1999/03/30/section-i-the-core-principles-of-journalism>. Accessed April 2016.
- Shorenstein Center on Media, Politics and Public Policy. *Understanding data journalism: Overview of resources, tools and topics*. <http://journalistsresource.org/reference/reporting/understanding-data-journalism-overview-tools-topics>. Accessed April 2016.
- Silver, N. *What the fox knows*. <http://fivethirtyeight.com/features/what-the-fox-knows>. Accessed August 2014.

Special Issues and Volumes

- Digital Journalism—Journalism in an Era of Big Data: Cases, concepts, and critiques*. v. 3/3 (2015).
- Social Science Computer Review – Citizenship, Social Media, and Big Data: Current and Future Research in the Social Sciences* (in press).
- The ANNALS of American of the American Academy of Political and Social Science – Toward Computational Social Science: Big Data in Digital Environments*. v. 659/1 (2015).

K

KDD

► [Data Discovery](#)

KDDM

► [Data Discovery](#)

Keycatching

► [Keystroke Capture](#)

Keylogger

► [Keystroke Capture](#)

Keystroke Capture

Gordon Alley-Young
Department of Communications and Performing
Arts, Kingsborough Community College, City
University of New York, New York, NY, USA

Synonyms

[Keycatching](#); [Keylogger](#); [Keystroke logger](#);
[Keystroke recorder](#)

Introduction

Keystroke capture (KC) tracks a computer or mobile device users' keyboard activity using hardware or software. KC is used by businesses to keep employees from misusing company technology, in families to monitor the use possible misuse of family computers, and by computer hackers who seek gain through secretly possessing an individual's personal information and account passwords. KC software can be purchased for use on a device or may be placed

maliciously without the user's knowledge through contact with untrusted websites or e-mail attachments. KC hardware can also be purchased and is disguised to look like computer cords and accessories. KC detection can be difficult because software and hardware are designed to avoid detection by anti-KC programs. KC can be avoided by using security software as well as through careful computing practices. KC affects individual computer users as well as small, medium, and large organizations internationally.

How Keystroke Capture (KC) Works

Keystroke capture (KC), also called keystroke logger, keylogger, keystroke recorder, and keycatching, tracks a computer or mobile device users' activities, including keyboard activity, using hardware or software. KC is knowingly employed by businesses to deter its employees from misusing company devices and also by families seeking to monitor the technology activities of vulnerable family members (e.g., teens, children). Romantic partners and spouses use KC to catch their significant others engaged in deception and/or infidelity. Computer hackers install KC onto unsuspecting users' devices in order to steal their personal data, website passwords, financial information, read their correspondence/online communication, to stalk/harass/intimidate users, and/or to sabotage organizations or individuals that hackers consider unethical. When used covertly to hurt and/or steal from others, KC is called malware, malicious software used to interfere with a device, and/or spyware, software used to steal information or to spy on someone.

KC software (e.g., WebWatcher, SpectorPro, Cell Phone Spy) is available for free and also for purchase, and it is usually downloaded onto the device where it either saves captured data onto the hard drive or sends it through networks/wirelessly to another device/website. KC hardware (e.g., KeyCobra, KeyGrabber, KeyGhost) may be an adaptor device into which a keyboard/mouse USB cord is plugged before it is inserted in to the computer or may look like an extension cable. Hardware can also be installed inside the

computer/keyboard. KC is placed on devices maliciously by hackers when computer and mobile device users visit websites, open e-mail attachments, or click links to files that are from untrusted sources. Individual technology users are frequently lured by untrusted sources and websites that offer free music files or pornography. KC's infiltrate organizations' computers when an employee is completing company business (i.e., financial transactions) on a device that he/she also uses to surf the Internet in their free time.

When a computer is infected with a malicious KC, it can be turned into what is called a zombie, a computer that is hijacked and used to spread KC malware/spyware to other unsuspecting individuals. A network of zombie computers that is controlled by someone other than the legitimate network administrator is called a botnet. In 2011, the FBI shut down the Coreflood botnet, a global KC operation affecting 2 million computers. This botnet spread KC software via an infected e-mail attachment and seemed to infect only computers using Microsoft Windows operating systems. The FBI seized the operators' computers and charged 13 "John Doe" defendants with wire fraud, bank fraud, and illegally intercepting electronic communication. Then in 2013 security firm SpiderLabs found 2 million passwords in the Netherlands stolen by the Pony botnet. While researching the Pony botnet, SpiderLabs discovered that it contained over a million and a half Twitter and Facebook passwords and over 300,000 Gmail and Yahoo e-mail passwords. Payroll management company ADP, with over 600,000 clients in 125 countries, was also hacked by this botnet.

The Scope of the Problem Internationally

In 2013 the Royal Canadian Mounted Police (RCMP) served White Falcon Communications with a warrant that alleged that the company was controlling an unknown number of computers known as the Citadel botnet (Vancouver Sun 2013). In addition to distributing KC malware/

spyware, the Citadel botnet also distributed spam and conducted network attacks that reaped over \$500 million dollars illegal profit affecting more than 5 million people globally (Vancouver Sun 2013). The Royal Bank of Canada and HSBC in Great Britain were among the banks attacked by the Citadel botnet (Vancouver Sun 2013). The operation is believed to have originated from Russia or Ukraine as many websites hosted by White Falcon Communications end in the .ru suffix (i.e., country code for Russia). Microsoft claims that the 1,400 botnets running Citadel malware/spyware were interrupted due to the RCMP action with the highest infection rates in Germany (Vancouver Sun 2013). Other countries affected were Thailand, Italy, India, Australia, the USA, and Canada. White Falcon owner Dmitry Glazyrin's voicemail claimed he was out of the country on business when the warrant was served (Vancouver Sun 2013).

Trojan horses allow others to access and install KC and other malware. Trojan horses can alter or destroy a computer and its files. One of the most infamous Trojan horses is called Zeus. Don Jackson, a senior security researcher with Dell SecureWorks and who has been widely interviewed, claims that Zeus is so successful because those behind it, seemingly in Russia, are well funded and technologically experienced, and this allows them to keep Zeus evolving into different variations (Button 2013). In 2012 Microsoft's Digital Crimes Unit with its partners disrupted a variation of Zeus botnets in Pennsylvania and Illinois responsible for an estimated 13 million infections globally. Another variation of Zeus called GameOver tracks computer users' every login and uses the information to lock them out and drain their bank accounts (Lyons 2014). In some instances GameOver works in concert with CryptoLocker. If GameOver finds that an individual has little in the bank then CryptoLocker will encrypt users' valuable personal and business files agreeing to release them only once a ransom is paid (Lyons 2014). Often ransoms must be paid in Bitcoin, Internet based and currently anonymous and difficult to track. Victims of CryptoLocker will often receive a request for a one Bitcoin ransom (estimated to be worth 400€/\$500USD)

to unlock the files on their personal computer that could include records for a small business, academic research, and/or family photographs (Lyons 2014).

KC is much more difficult to achieve on a smartphone as most operating systems operate only one application at a time, but it is not impossible. As an experiment Dr. Hao Chen, an Associate Professor in the Department of Computer Science at the University of California, Davis, with an interest in security research created a KC software that operates using smartphone motion data. When tested, Chen's application correctly guessed more than 70% of the keystrokes on a virtual numerical keypad though he asserts that it would probably be less accurate on an alphanumeric keypad (Aron 2011). Point-of-sale (POS) data, gathered when a credit card purchase is made in a retail store or restaurant, is also vulnerable to KC software (Beierly 2010). In 2009 seven Louisiana restaurant companies (i.e., Crawfish Town USA Inc., Don's Seafood & Steak House Inc., Mansy Enterprises LLC, Mel's Diner Part II Inc., Sammy's LLC, Sammy's of Zachary LLC, and B.S. & J. Enterprises Inc.) sued Radiant Systems Inc., a POS system maker, and Computer World Inc., a POS equipment distributor, charging that the vendors did not secure the Radiant POS systems. The customers were then defrauded by KC software, and restaurant owners incurred financial costs related to this data capture. Similarly, Patco Construction Company, Inc. sued People's United Bank for failing to implement sufficient security measures to detect and address suspicious transactions due to KC. The company finally settled for \$345,000, the cost that was stolen plus interest.

Teenage computer hackers, so-called *hactivists* (people who protest ideologically by hacking computers), and governments under the auspices of cyber espionage engage in KC activities, but cyber criminals attain the most notoriety. Cyber criminals are as effective as they are evasive due to the organization of their criminal gangs. After taking money from bank accounts via KC, many cyber criminals send the payments to a series of money mules. Money mules are sometimes unwitting participants in fraud who are recruited via the Internet with promises of money for

working online. The mules are then instructed to wire the money to accounts in Russia and China (Krebs 2009). Mules have no face-to-face contact with the heads of KC operations so it can be difficult to secure prosecutions, though several notable cyber criminals have been identified, charged, and/or arrested. In late 2013 the RCMP secured a warrant for Dmitry Glazyrin, the apparent operator of a botnet who left Canada before the warrant could be served. Then in early 2014, Russian SpyEye creator Aleksandr Panin was arrested for cyber crime (IMD 2014). Also, Estonian Vladimir Tsastsin, the cyber criminal who created DNSChanger and became rich of online advertising fraud and KC by infecting millions of computers. Finnish Internet security expert Mikko Hermann Hyppönen claimed that Tsastsin owned 159 Estonian properties when he was arrested in 2011 (IMD 2014). Tsastsin was released 10 months after his arrest due to insufficient proof. As of 2014 Tsastsin has been extradited to the US for prosecution (IMD 2014). Also in 2014 the US Department of Justice (DOJ) filed papers accusing a Russian Evgeniy Mikhailovich Bogachev of leading the gang behind GameOver Zeus. The DOJ claims GameOver Zeus caused \$100 million in losses from individuals and large organizations.

Suspected Eastern European malware/spyware oligarchs have received ample media attention for perpetuating KC via botnets and Trojan horses while other perpetrators have taken the public by surprise. In 2011 critics accused software company Carrier IQ of placing KC and geographical position spyware in millions of users' Android devices (International Business Times 2011). The harshest of critics have alleged illegal wiretapping on the part of the company while Carrier IQ has rebutted that what was identified as spyware is actually diagnostic software that provides network improvement data (International Business Times 2011). Further the company stated that the data was both encrypted and secured and not sold to third parties. In January 2014, 11 students were expelled from Corona del Mar High School in California's affluent Orange County for allegedly using KC to cheat for several years with the help of tutor Timothy Lai. Police

report being unable to find Lai, a former resident of Irvine, CA, since the allegations surfaced in December 2013. The students are accused of placing KC hardware onto teachers' computers to get passwords to improve their grades and steal exams. All 11 students signed expulsion agreements in January 2014 that whereby they abandoned their right to appeal their expulsions in exchange for being able to transfer to other schools in the district. Subsequently, five of the students' families sued the district for denying the students the right to appeal and/or claiming tutor Lai committed the KC crimes. By the end of March, the school district had spent almost \$45,000 in legal fees.

When large organizations are hacked via KC, the news is reported widely. For instance, Visa found KC software being able to transmit card data to a fixed e-mail or IP address where hackers could retrieve it. Here the hackers attached KC to a POS system. Similarly KC was used to capture the keystrokes of pilots flying the US military's Predator and Reaper drones that have been used in Afghanistan (Shachtman 2011). Military officials were unsure whether the KC software was already built into the drones was the work of a hacker (Shachtman 2011). Finally, Kaspersky Labs has publicized how it is possible to get control of BMW's Connected Drive system via KC and other malware, and this gain control of a luxury car that uses this Internet-based system.

Research by Internet security firm Symantec shows that many small and medium-sized businesses believe that malware/spyware is a problem for large organizations (e.g., Visa, the US military). However, since 2010 the company notes that 40% of all companies attacked have fewer than 500 employees while only 28% of attacks target large organizations. A case in point is a 2012–2013 attack on a California escrow firm, Efficient Services Escrow Group of Huntington Beach, CA, that had one location and nine employees. Using KC malware/spyware, the hackers drained the company of \$1.5 million dollars in three transactions wired to bank accounts in China and Russia. Subsequently, \$432,215 sent to a Moscow Bank was recovered, while the \$1.1 million sent to China was never recouped. The

loss was enough to shutter the business's one office and put its nine employees out of work.

Though popular in European computer circles, the relatively low-profile Chaos Computer Club learned that German state police were using KC malware/spyware as well as saving screenshots and activating the cameras/microphones of club members (Kulish and Homola 2014). News of the police's actions led the German justice minister to call for stricter privacy rules (Kulish and Homola 2014). This call echoes a 2006 commission report to the EU Parliament that calls for strengthening the regulatory framework for electronic communications. KC is a pressing concern in the US for as of 2014, 18 states and one territory (i.e., Alaska, Arizona, Arkansas, California, Georgia, Illinois, Indiana, Iowa, Louisiana, Nevada, New Hampshire, Pennsylvania, Rhode Island, Texas, Utah, Virginia, Washington, Wyoming, Puerto Rico) all have anti-spyware laws on the books (NCSL 2015).

Tackling the Problem

The problem of malicious KC can be addressed through software interventions and changes in computer users' behaviors, especially when online. Business travelers may be at a greater risk for losses if they log onto financial accounts using hotel business centers as these high-traffic areas provide ample opportunities to hackers (Credit Union Times 2014). Many Internet security experts recommend not using public wireless networks where of KC spyware thrives. Experts at Dell also recommend that banks have separate computers dedicated only to banking transactions with no emailing or web browsing.

Individuals without the resources to devote one computer to financial transactions can, experts argue, protect themselves from KC through changing several computer behaviors. First, individuals should change their online banking passwords regularly. Second, they should not use the same password for multiple accounts or use common words or phrases. Third is checking one's bank account on a regular basis for unauthorized transfers. Finally, it is important to log off of

banking websites when finished with them and to never click on third-party advertisements that post to online banking sites and take you to a new website upon clicking.

Configurations of one's computer features, programs, and software are also urged to thwart KC. This includes removing remote access (i.e., accessing one's work computer from home) configurations when they are not needed in addition to using a strong firewall (Beierly 2010). Users need to continually check their devices for unfamiliar hardware attached to mice or keyboards as well as check the listings of installed software (Adhikary et al. 2012; Beierly 2010). Many financial organizations are opting for virtual keypads and virtual mice, especially for online transactions (Kumar 2009). Under this configuration instead of typing a password and username on the keyboard using number and letter keys, the user scrolls through numbers and letters using the cursors' virtual keyboard. Always use the online virtual keyboard for your banking password to avoid the risk of keystrokes being logged when available.

Conclusion

Having anti-KC/malware/spyware alone does not guarantee protection, but experts agree that it is an important component of an overall security strategy. Anti-KC programs include SpyShelter Stop-Logger, Zemana AntiLogger, KeyScrambler Premium, Keylogger Detector, and GuardedID Premium. Some computer experts claim that PC's are more susceptible to KC malware/spyware than are Mac's as KC malwares/spywares are often reported to exploit holes in PC's operating systems, but new wisdom suggests that all devices can be vulnerable especially when programs and plug-ins are added to devices. Don Jackson, a senior security researcher with Dell SecureWorks, argues that one of the most effective methods for preventing online business fraud, the air-gap technique, is not widely utilized despite being around since 2005. The air-gap technique creates a unique verification code that is transmitted as a digital token, text message, or other device not connected

to the online account device, so the client can read and then key in the code as a signature for each transaction over a certain amount. Alternately in 2014 Israeli researchers presented research on a technique to hack an air-gap network using just a cellphone.

Cross-References

- [Cyber Espionage](#)
- [Data Brokers and Data Services](#)
- [Industrial and Commercial Bank of China](#)

Further Reading

- Adhikary, N., Shrivastava, R., Kumar, A., Verma, S., Bag, M., & Singh, V. (2012). Battering keyloggers and screen recording software by fabricating passwords. *International Journal of Computer Network & Information Security*, 4(5), 13–21.
- Aron, J. (2011). Smartphone jiggles reveal your private data. *New Scientist*, 211(2825), 21.
- Beierly, I. (2010). *They'll be watching you*. Retrieved from http://www.hospitalityupgrade.com/files/File_Articles/HUSum10_Beierly_Keylogging.pdf.
- Button, K. (2013). *Wire and online banking fraud continues to spike for businesses*. Retrieved from http://www.americanbanker.com/issues/178_194/wire-and-online-banking-fraud-continues-to-spike-for-businesses-1062666-1.html.
- Credit Union Times. (2014). Hotel business centers hacked. *Credit Union Times*, 25(29), 11.
- IMD: International Institute for Management Development. (2014). *Cybercrime buster speaks at IMD*. Retrieved from <http://www.imd.org/news/Cybercrime-buster-speaks-at-IMD.cfm>.
- International Business Times. (2011). *Carrier iq spyware: Company's Android app logging the keystrokes of millions*. Retrieved from <http://www.ibtimes.com/carrier-iq-spyware-companys-android-app-logs-keystrokes-millions-video-377244>.
- Krebs, B. (2009). *Data breach highlights role of 'money mules'*. Retrieved from http://voices.washingtonpost.com/securityfix/2009/09/money_mules_carry_loot_for_org.html.
- Kulish, N., & Homola, V. (2014). *Germans condemn police use of spyware*. Retrieved from http://www.nytimes.com/2011/10/15/world/europe/uproar-in-germany-on-police-use-of-surveillance-software.html?_r=0.
- Kumar, S. (2009). Handling malicious hackers & assessing risk in real time. *Siliconindia*, 12(4), 32–33.
- Lyons, K. (2014). *Is your computer already infected with dangerous Gameover Zeus software? Virus could be lying dormant in thousands of Australian computers*. Retrieved from <http://www.dailymail.co.uk/news/article-2648038/Gameover-Zeus-lying-dormant-thousands-Australian-computers-without-knowing.html#ixzz3AmHLKIZ9>.
- NCSL: National Conference of State Legislatures. (2015). *State spyware laws*. Retrieved from <http://www.ncsl.org/research/telecommunications-and-information-technology/state-spyware-laws.aspx>.
- Shachtman, N. (2011). *Exclusive: Computer virus hits US drone fleet*. Retrieved from <http://www.wired.com/2011/10/virus-hits-drone-fleet/>.
- Vancouver Sun. (2013). *Police seize computers linked to large cybercrime operation: Malware Responsible for over \$500 million in losses has affected more than five million people globally*. Retrieved from <http://www.vancouversun.com/news/Police+seize+computers+linked+large+cybercrime+operation/8881243/story.html#ixzz3Ale1G13s>.

Keystroke Logger

- [Keystroke Capture](#)

Keystroke Recorder

- [Keystroke Capture](#)

Key-Value-Based Database

- [NoSQL \(Not Structured Query Language\)](#)

Knowledge Discovery

- [Data Discovery](#)

Knowledge Graph

- [Ontologies](#)

Knowledge Hierarchy

► Data-Information-Knowledge-Action Model

Knowledge Management

Magdalena Bielenia-Grajewska
Division of Maritime Economy, Department of
Maritime Transport and Seaborne Trade,
University of Gdansk, Gdansk, Poland
Intercultural Communication and
Neurolinguistics Laboratory, Department of
Translation Studies, University of Gdansk,
Gdansk, Poland

There are different definitions of knowledge management. As Gorelick et al. (2004, p. 4) state, “knowledge management is a vehicle to systematically and routinely help individuals, groups, teams, and organizations to: learn what the individual knows; learn what others know (e.g. individuals and teams); learn what the organization knows; learn what you need to learn; organize and disseminate these learnings effectively and simply; apply these learnings to new endeavours”. Knowledge can also be defined by juxtaposing it with another phenomenon, being relatively close to it. As Foray (2006, p. 4) claims, “in my conception, knowledge has something more than information: knowledge-in whatever field- empowers its possessors with the capacity for intellectual or physical action. What I mean by knowledge is fundamentally a matter of cognitive ability. Information, on the other hand, takes the shape of structured and formatted data that remain passive and inert until used by those with the knowledge needed to interpret and process them.” He adds that “therefore, the reproduction of knowledge and the reproduction of information are clearly different phenomena. While one takes place through learning, the other takes place simply through duplication. Mobilization of a cognitive resource is always necessary for the reproduction of knowledge, while information can be reproduced by a photocopy machine” (Foray 2006, p. 4). In short,

knowledge management (KM) can be defined as a set of tools and methods connected with organizing knowledge. It encompasses such activities as creating, encoding, systematizing, distributing and acquiring knowledge. There are number of reasons why knowledge management is very crucial in modern times. First of all, it should be mentioned that the twenty-first century can be characterized by the large amount of data that modern people are surrounded by. Secondly, many spheres of modern life depend on knowledge flows; information societies demand not only the access to knowledge but also its effective management. Thirdly, technological advancements facilitate the effectiveness connected with different stages of knowledge management. Thus, the need to manage knowledge has become more important nowadays than it was in other centuries. Knowledge management is classified by taking into account both the process and subject approach. However, the processual perspective reflects the changing nature of knowledge that has to constantly adapt to new conditions of the environment and expectations of the target audience. Thus, KM is studied by taking into account the processes accompanying creating, codifying, disseminating as well as teaching and learning. Apart from processes, KM should also be investigated from the prism of different types of knowledge involved in knowledge management. As far as other features of knowledge management are concerned, Jermielniak (2012) stresses that knowledge is a primary resource that allows other resources to be created and acquired. Moreover, in the process of using, knowledge does not use up but grows continually.

Types of Knowledge

Knowledge can be classified by taking into account different factors. The famous division is the one by Nonaka and Konno (1998) who discuss the concepts of tacit and explicit knowledge. “Explicit knowledge can be expressed in words and numbers and shared in the form of data, scientific formulae, specifications, manuals, and the like. Tacit knowledge is highly personal and hard to formalize, making it difficult to communicate or share with others” (Nonaka and Konno

1998, p. 42). Another way is to look at knowledge management through the prism of knowledge architects. The first notion that can be taken into account is the level of professionalism among information creators. Thus, such types of knowledge can be distinguished as professional/expert knowledge and laymen knowledge. *Professional/expert knowledge* is connected with knowledge that can be acquired exclusively by vocational schooling, professional experience, and/or specialized training. On the other hand, *laymen knowledge* is associated with the knowledge on the topic possessed by an average human being, resulting from one's experience with using, e.g., a given device or information gained from others. Knowledge can also be categorized by taking into account the acceptable level of information disclosure; consequently, open and closed knowledge can be categorized. *Open knowledge* is available freely to everybody, whereas *closed knowledge* is directed at a selected group of users. An example of open knowledge is an article published online in the open-access journal, whereas the same article published in the journal with subscription belongs to closed knowledge. Knowledge can also be classified by taking into account the notion of tangibility. *Tangible knowledge* is the type of knowledge that can be easily perceived and measured (e.g., by points and marks). On the other hand, *intangible knowledge* encompasses knowledge that cannot be easily perceived and managed. Knowledge can also be classified by analyzing the channel used for creating and disseminating knowledge. The first division concerns the type of sense used for knowledge processing- the most general division includes the classification into verbal and non verbal knowledge. *Verbal knowledge* encompasses knowledge produced and disseminated in a verbal way, by relying on a commonly known linguistic system of communication. Verbal knowledge includes, e.g., words and phrases characteristic of a given language and culture. Verbal knowledge can also be subcategorized by observing, e.g., the length of information. Thus, the micro approach encompasses morphemes, words, and phrases, the meso dimension focuses on texts, whereas the macro dimension concerns,

e.g., corporate or national linguistic policies. *Non-verbal knowledge* encapsulates other than verbal types of knowledge. For example, *auditory knowledge* encompasses elements of knowledge disseminated through the audio channel; it is represented in jingles and songs. *Olfactory knowledge* includes knowledge gained by the sense of smell and it concerns, e.g., the flavors connected with regional festivities. Another type of knowledge is *tactile knowledge*, being the type of knowledge acquired through the physical experience of touching objects. The advancement in modern technology has also led to the classification of online knowledge and offline knowledge. *Online knowledge* is the type of knowledge created and made available on the Internet, whereas *offline knowledge* encompasses the knowledge made and published outside the web.

Knowledge Transfer

Knowledge transfer or knowledge flow can be briefly defined as moving the knowledge from one person/organization to another one. Teece (2001) divides knowledge transfer into internal and external knowledge transfer. *Internal transfer* takes place within an organization, e.g., between workers, whereas *external transfer* takes place from one company to another. The latter one includes technology transfer and intellectual property rights. As Zizzo (2005) claims, transfer of knowledge can be vertical or horizontal. *Vertical transfer of knowledge* is connected with using rules and characteristics in similar situations, whereas *horizontal transfer of knowledge* is represented in direct and context-dependant adaptation of problem to similar one.

Factors Determining Knowledge Management

One of the key factors determining knowledge management is language. The first linguistic issue shaping KM is the opportunity to access information, taking into account the language used to create and disseminate knowledge. Thus, the lack of linguistic skills in a given language may lead to the limited or no access to required data. For example, knowledge created and

disseminated in English can be reached only by the users of the lingua franca. To meet the growing demand of knowledge among linguistically diverse users, translation is one of the methods directed at effective knowledge management. Another approach to organizational knowledge management can be observed in many international companies; they adopt a corporate linguistic policy that regulates linguistic matters in business entities by analyzing both corporate and individual linguistic needs. Apart from languages understood in the broad sense, that is, as the linguistic repertoire used by a nation or a large cultural group, they may also be studied by taking into account dialects or professional genres. As far as knowledge management is concerned, attention is focused on making knowledge created within a small linguistic community relatively accessible to all interested stakeholders. Analyzing the example of corporations, knowledge produced in, e.g., professional discourse of accountants should be comprehensible to the representative of other professions during corporate meetings. Linguistic knowledge management also concerns the selection of linguistic tools to manage knowledge effectively. An example of the mentioned linguistic approach is to select the verbs or adjectives that are supposed to attract the readers to knowledge, make the content reliable and informative, as well as invoke certain reactions. In addition, the discussion on the linguistic tools should also encompass the role of literal and non-literal language in KM. As Bielenia-Grajewska (2014, 2015, 2018) stresses, the nonliteral dimension encompasses the figurative tools used in discourse, such as idioms, puns, similes, and metaphors. Taking the example of metaphors, they serve different functions in effective knowledge management. First of all, metaphors facilitate the understanding of complex and novel information. Using a familiar domain to explain a new concept turns out to be very effective. Secondly, metaphors serve as an economical way of providing information. Instead of using long phrases to facilitate explanation of novel approaches in knowledge, a metaphor relying on a well-known domain makes the concept comprehensible. The next important factor of knowledge management is technology; technological advancements

accompany all stages of KM. For example, technology facilitates the development of knowledge as well as its subsequent presentation in online and offline informational outlets and special databases. Knowledge management is also determined by individual and social factors. Taking into account the personal dimension, the attitude to knowledge management is shaped by such factors as gender, age, profession and interest in innovation. Knowledge management also depends on group features, namely, how a given community perceives the importance of knowledge management. Knowledge management also depends on political and economic situation of a country. Thus, it should be stated that there are different factors of micro, meso, and macro nature that determine the way knowledge is created, disseminated, and stored.

Fantino and Stolarz-Fantino (2005) discuss the role of different types of context, such as spatial context, temporal context, historical context and linguistic context in understanding one's behavior. They also stress that context can be understood in different ways. "Context refers to many aspects of our environment that play an important role in determining our behavior. For example, in the laboratory, the term context may be used to refer to any of the following: background stimuli that affect the degree of conditioning to foreground stimuli; historical events that affect subjects' appreciation of contemporary stimuli; rules or superordinate stimuli that stipulate the correct response to a target stimulus in a situation. In the laboratory it is simple to demonstrate *stimulus control*, by which we mean that a behavior will be maintained in the presence of one stimulus (or context) but not in another. The more similar two contexts (or stimulus configurations), the more likely a behavior acquired in the presence of one context is likely to transfer to (occur in the presence of) the other context" (2005, p. 28). Nonaka and Konno (1998) discuss that knowledge is embedded in *ba* (shared spaces). *Ba* is a concept coined by the Japanese philosopher Kitaro Nishida to denote a shared space for emerging relationships. The space can be of different character: physical (e.g. office, home), virtual (e.g. online chat), mental (shared experience), or the combination of the mentioned features. *Ba* is

perceived as a platform for facilitating individual and collective knowledge.

Knowledge Management and Methodology

Researchers rely on different methods investigating KM. For example, the selection of tools depends on the nature of knowledgeable elements. Methodology can be classified, e.g., by taking into account the type of stimuli, such as verbal and nonverbal elements. Thus, audio elements, such as sounds, songs, or jingles, require different methods of data management than olfactory data. Taking into account the magnitude of factors determining KM, there are certain methodologies that prove to be used more often than the other ones. One of the methods applied to study KM is the network approach. Within the network perspectives, Bielenia-Grajewska (2011) highlights the one called actor-network analysis which, stressing the importance of both living and nonliving entities in the way a given person, thing, or organization performs, turns out to be useful in KM. Applying the ANT approach, it is possible to highlight the role of things, such as telephones or computers, in transmitting and storing knowledge. In addition, it can be researched how, e.g., the technological defects of machines influence the effectiveness of knowledge management. Another network technique – social network analysis – concentrates on the relations between individuals. This approach may provide information how data is distributed among network members and how the types of nodes and ties determine the way KM is handled. Apart from interdisciplinary approaches, KM may also rely on various disciplines. It should also be stated that the methods used in linguistics may support the research on knowledge management. For example, critical discourse analysis facilitates the understanding of verbs, nouns, adjectives, or numerals in managing knowledge. CDA may help to create a text that will be understandable by a relatively large group of people; verbal and pictorial tools of communication are studied to show how they separately as well as together

determine knowledge management on a given topic. In addition, CDA offers the option to study how knowledge management changes depending on the type of situations. For example, risky conditions demand other communication tools than the coverage of leisure activities. Another discipline that facilitates the research on knowledge management is neuroscience that offers a plethora of methods to investigate the way knowledge is perceived and understood. For example, modern neuroscientific apparatus makes it possible to study the effectiveness of knowledge. As Bielenia-Grajewska (2013) stresses, it is visible in the application of neuroscientific tools in modern management. One of the possible tools used in neuroscientific investigations is fMRI. *Functional magnetic resonance imaging* is a technique that uses the advancements of magnetic resonance to research brain performance. The investigation concerns mainly two stages. The first part is devoted to taking the anatomical scans of the subject's brain when the person lies still in a scanner. The next stage concerns the active involvement of a subject in some activity (e.g., choosing a word, a phrase, or a picture). The apparatus measures the BOLD signal (blood oxygen level dependent) that shows which parts of the brain are active. Such experiments facilitate effective knowledge management since such experiments show which pieces of information are easier understood. It should be mentioned that also the investigations on other parts of the body may provide information on how knowledge is understood. The emotional response to knowledge management can be investigated by analyzing the way face muscles respond to a given stimulus. *Facial electromyography* (fEMG) measures the face muscles nerve (e.g., the zygomatic major muscle) when the subject is shown a stimulus. In addition, the emotional attitude of the subject to the presented knowledge can be checked by observing the electrodermal reactions, using the technique called *galvanic skin response*. Researchers may also observe the heart rate or blood pressure to check the reaction of the subject to a given stimulus. In addition, knowledge management can be researched from both qualitative and quantitative perspectives. As far as the

quantitative focus is concerned, knowledge management can be supported by statistical tools that organize big data. It should also be stated that the growing role of the Internet in modern life has led to the interest in online and offline approaches of knowledge management. Knowledge management uses different tools to disseminate knowledge in a quick and effective way. One of the ways is to use stories in KM. As Gorelick et al. (2004) state, real stories based on one's experience become codified and become available in the form of knowledge assets. Among different tools, Probst (2002) discusses the role of case-writing for knowledge management. First, he mentions that case writing is used as a teaching tool in e.g. MBA studies since it offers students learning new knowledge from real-life situations. Secondly, the narrative style of case writing offers discussion and reflection on issues presented in cases. Thirdly, they are an effective tool for increasing the skills and knowledge of managers. He suggests that companies should write the cases about their situations that show how experience and knowledge were acquired through the period of time. In the case of collective case writing, learning is fostered in a spiral way from the individual, through the group, to the corporate level.

Knowledge and Big Data

The place of knowledge management in big data is often discussed by taking into account the novelties in the sphere of dealing with information. Knowledge nowadays can be extracted from different types of data, namely, structured data and unstructured data. *Structured data* is well-organized and can be found in different databases. It may include names, telephones, addresses, among others. *Unstructured data*, on the other hand, is not often to be found in databases and is not as searchable as structured data is. It includes material of different nature, such as written, video, or audio ones, including websites, texts, emails, and conversations. In the area of big data, information exists in different types and in different quantities that can be extracted by both humans and machines. As Neef (2015) discusses, two

concepts are associated with big data. *Social intelligence* is connected with monitoring social media and paying attention to data connected with likes, dislikes, sentiment data, and brand names. *Social analytics* comprises the tools applied to analyzed data, connected with what users discuss (share), what their opinion about these things is (engagement), and the way they discuss them (reach).

Future of Knowledge Management

It can be predicted that the increase in the amount of knowledge should be supported with more advanced tools that will enable not only to acquire extensive data but also to use and store it. Thus, it can be estimated that the future knowledge management will depend even more on the advancements of modern technology. One example of using the improvements in other domains of science is the application of neuroscientific expertise in the field of knowledge management as well as statistical methods aimed at analyzing large quantities of data. In addition, since big data are becoming more and more important in the reality of the twenty-first century, knowledge management has to rely on diverse and complicated tools that will facilitate the creation and dissemination of data. Consequently, the interrelation between KM and other disciplines is supposed to be growing in the coming years.

Cross-References

- [Information Society](#)
- [Social Media](#)
- [Social Network Analysis](#)
- [Statistics](#)

Further Reading

- Bielenia-Grajewska, M. (2011). A potential application of actor network theory in organizational studies: The company as an ecosystem and its power relations from the ANT perspective. In A. Tatnall (Ed.), *Actor-network theory and technology innovation:*

- Advancement and new concepts*. Hershey: Information Science Reference.
- Bielenia-Grajewska, M. (2013). International neuromanagement. In D. Tsang, H. H. Kazeroony, & G. Ellis (Eds.), *The Routledge companion to international management education*. Abingdon: Routledge.
- Bielenia-Grajewska, M. (2014). CSR online communication: The metaphorical dimension of CSR discourse in the food industry. In R. Tench, W. Sun, & B. Jones (Eds.), *Communicating corporate social responsibility: Perspectives and practice (critical studies on corporate responsibility, governance and sustainability, volume 6)*. Bingley: Emerald Group Publishing Limited.
- Bielenia-Grajewska, M. (2015). The role of figurative language in knowledge management. Knowledge encoding and decoding from the metaphorical perspective. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology*. Hershey: IGI Publishing.
- Bielenia-Grajewska, M. (2018). Knowledge management from the metaphorical perspective. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (4th ed.). Hershey: IGI Publishing.
- Fantino, E., & Stolarz-Fantino, S. (2005). Context and its effect on transfer. In D. J. Zizzo (Ed.), *Transfer of knowledge in economic decision making*. Basingstoke: Palgrave Macmillan.
- Foray, D. (2006). *The economics of knowledge*. Cambridge, MA: The MIT Press.
- Gorelick, C., Milton, N., & April, K. (2004). *Performance through learning. Knowledge Management in Practice*. Oxford: Elsevier.
- Jemielniak, D. (2012). Zarządzanie wiedzą. Podstawowe pojęcia. In D. Jemielniak & A. K. Koźmiński (Eds.), *Zarządzanie wiedzą*. Warszawa: Oficyna Wolters Kluwer.
- Neef, D. (2015). *Digital exhaust: What everyone should know about big data, digitization and digitally driven innovation*. Upper Saddle River: Pearson Education.
- Nonaka, I., & Konno, N. (1998). The concept of Ba. Building a foundation for knowledge creation. *California Management Review*, 40(3):40–54.
- Probst, G. J. B. (2002). Putting knowledge to work: Case-writing as a knowledge management and organizational learning tool. In T. H. Davenport & G. J. B. Probst (Eds.), *Knowledge management case book*. Erlangen: Publicis Corporate Publishing and John Wiley & Sons.
- Teece, D. J. (2001). Strategies for managing knowledge assets: The role of firm structure and industrial context. In I. Nonaka & D. J. Teece (Eds.), *In managing industrial knowledge: Creation, transfer and utilization*. London: SAGE Publications.
- Zizzo, D. J. (2005). Transfer of knowledge and the similarity function in economic decision-making. In D. J. Zizzo (Ed.), *Transfer of knowledge in economic decision making*. Basingstoke: Palgrave Macmillan.

Knowledge Pyramid

► Data-Information-Knowledge-Action Model

LexisNexis

Jennifer J. Summary-Smith
Florida SouthWestern State College, Fort Myers,
FL, USA
Culver-Stockton College, Canton, MO, USA

As stated on its website, LexisNexis is a leading global provider of content-enabled workflow solutions. This corporation provides data and solutions for professionals in areas such as the academia, accounting, corporate world, government, law enforcement, legal, and risk management. LexisNexis is a subscription-based service, with two data centers located in Springfield and Miamisburg, Ohio. The centers are among the largest complexes of their kind in the United States, providing LexisNexis with “one of the most complete comprehensive collections of online information in the world.”

Data Centers

The LexisNexis data centers hold network servers, software, and telecommunication equipment, which is a vital component of the entire range of LexisNexis products and services. The data centers service the LexisNexis Group Inc. providing assistance for application development, certification and administrative services,

and testing. The entire complex serves its Reed Elsevier sister companies while also providing LexisNexis customers with the following: backup services, data hosting, and online services. LexisNexis opened its first remote data center and development facility in Springfield, Ohio, in 2004, which hosts new product development. Both data centers function as a backup and recovery facility for each other.

According to the LexisNexis’ website, its customers use services that span multiple servers and operating systems. For example, when a subscriber submits a search request, the systems explore and sift through massive amounts of information. The answer set is typically returned to the customer within 6–10 s, resulting in a 99.99% average for reliability and availability of the search. This service is accessible to five million subscribers, with nearly five billion documents of source information available online and stored in the Miamisburg facility. The online services also provide access to externally hosted data from the Delaware Secretary of State, Dun & Bradstreet Business Reports, Historical Quote, and Real-Time Quote. Given that a large incentive for data center services is to provide expansion capacity for all future hosting opportunities, this has led to an increase in the percentage of total revenue for Reed Elsevier. Currently, the Miamisburg data center supports over two billion dollars in online revenue for Reed Elsevier.

Mainframe Servers

There are over 100 servers housed in the Springfield center, managing over 100 terabytes of data storage. As for the Miamisburg location, this complex holds 11 huge mainframe servers, running 34 multiple virtual storage (MVS) operating system images. The center also has 300 midrange Unix servers and almost 1,000 multiprocessor NT servers. They provide a wide range of computer services including patent images to customers, preeminent US case law citation systems, a hosting channel data for Reed Elsevier, and computing resources for the LexisNexis enterprise. As the company states, its processors have access to over 500 terabytes (or one trillion characters) of data storage capacity.

Telecommunications

LexisNexis has developed a large telecommunications network, permitting the corporation to support its data collection requirements while also serving its customers. As noted on its website, subscribers to the LexisNexis Group have a search rate of one billion times annually. LexisNexis also provides bridges and routers and maintains firewalls, high-speed lines, modems, and multiplexors, providing an exceptional degree of connectivity.

Physical Dimensions of the Miamisburg Data Center

LexisNexis Group has hardware, software, electrical, and mechanical systems housed in a 73,000 ft² data center hub. Its sister complex, located in Springfield, comprises a total of 80,000 ft². In these facilities, the data center hardware, software, electrical, and mechanical systems have multiple levels of redundancy, in the event that a single component fails, ensuring uninterrupted service. The company's website states that its systems are maintained and tested on a regular basis to ensure they perform correctly in case of an emergency. The LexisNexis Group also holds and stores copies of critical data off-

site. Multiple times a year, emergency business resumption plans are tested. Furthermore, the data center has system management services 365 days a year and 24 h a day provided by skilled operations engineers and staff. If needed, there are additional specialists on site, or on call, to provide the best support to customers. According to its website, LexisNexis invests a great deal in protection architecture to prevent hacking attempts, viruses, and worms. In addition, the company also has third-party contractors which conduct security studies.

Security Breach

In 2013, Byron Acohido reported that a hacking group hit three major data brokerage companies. LexisNexis, Dun & Bradstreet, and Kroll Background America are companies that stockpile and sell sensitive data. The group that hacked these data brokerage companies specialized in obtaining and selling social security numbers. The security breach was disclosed by a cybersecurity blogger Brian Kebs. He stated that the website ssndob.ms (SSNDOB), their acronym stands for social security number and date of birth, markets itself on underground cybercrime forums, offering services to its customers who want to look up social security numbers, birthdays, and other data on any US resident. LexisNexis found an unauthorized program called nbc.exe on its two systems listed in the botnet interface network located in Atlanta, Georgia. The program was placed as far back as April 2013, compromising their security for at least 5 months.

LexisNexis Group Expansion

As of July 2014, LexisNexis Risk Solutions expanded its healthcare solutions to the life science marketplace. In an article by Amanda Hall, she notes that an internal analysis revealed that 40% of the customer files have missing or inaccurate information in a typical life science company. LexisNexis Risk Solutions has leveraged its leading databases, reducing costs, improving

effectiveness, and strengthening identity transparency. LexisNexis is able to deliver data to over 6.5 million healthcare providers in the United States. This will benefit life science companies allowing them to tailor their marketing and sales strategies, to identify the correct providers to pursue. The LexisNexis databases are more efficient, which will help health science organizations gain compliance with federal and state laws.

Following the healthcare solutions announcement, Elisa Rodgers writes that Reed Technology and Information Services, Inc., a LexisNexis company, acquired PatentCore. PatentCore is an innovator of patent data analytics. PatentAdvisor is a user-friendly suite, delivering information to assist with a more effective patent prosecution and management. Its web-based patent analytic tools will help IP-driven companies and law firms by making patent prosecution a more strategic and probable process.

The future of the LexisNexis Group should include more acquisitions, expansion, and increased capabilities for the company. According to its website, the markets for their companies have grown over the last three decades, servicing professionals in academic institutes, corporations, governments, and business people. LexisNexis Group provides critical information, in easy-to-use electronic products, to the benefit of subscribed customers. The company has a long history of fulfilling its mission statement “to enable its customers to spend less time searching for critical information and more time using LexisNexis knowledge and management tools to guide critical decisions.” For more than a century, legal professionals have trusted the LexisNexis Group. It appears that the company will continue to maintain this status and remain one of the leading providers in the data brokerage marketplace.

Cross-References

- [American Bar Association](#)
- [Big Data Quality](#)
- [Data Brokers and Data Services](#)
- [Data Center](#)
- [Ethical and Legal Issues](#)

Further Reading

- Acohido, B. *LexisNexis, Dunn & Bradstreet, Kroll Hacked*. <http://www.usatoday.com/story/cybertruth/2013/09/26/lexisnexis-dunn-bradstreet-altegrity-hacked/2878769/>. Accessed July 2014.
- Hall, A. *LexisNexis verified data on more than 6.5 million providers strengthens identity transparency and reduces costs for life science organizations*. <http://www.benzinga.com/pressreleases/14/07/b4674537/lexisnexis-verified-data-on-more-than-6-5-million-providers-strenghtens>. Accessed July 2014.
- Krebs, B. *Data broker giants hacked by ID theft service*. <http://krebsonsecurity.com/2013/09/data-broker-giants-hacked-by-id-theft-service/>. Accessed July 2014.
- LexisNexis. <http://www.lexisnexis.com>. Accessed July 2014.
- Rodgers, E. *Adding multimedia reed tech strengthens line of LexisNexis intellectual property solutions by acquiring PatentCore, an innovator in patent data analytics*. <http://in.reuters.com/article/2014/07/08/supp-pa-reed-technology-idUSnBw015873a+100+BSW20140708>. Accessed July 2014.

Lightnet

- [Surface Web vs Deep Web vs Dark Web](#)

Link Prediction in Networks

Anamaria Berea

Department of Computational and Data Sciences,
George Mason University, Fairfax, VA, USA
Center for Complexity in Business, University of
Maryland, College Park, MD, USA

Link Prediction is an important methodology in social network analysis that aims to predict existing links between the nodes of a network when there is incomplete or partial information about the network. Link Prediction is also a very important method for assessing the development and evolution of dynamic networks. While link prediction is not a method only specific to big data, as it can be used with smaller datasets as well, its importance for big data arises from the complexity of large networks with varied

topologies and the importance of pattern identification that is specific only for large, complex datasets (Wei et al. 2017).

There have been developed a number of algorithms that are predicting the missing information from networks or reconstruct networks. The aim of these algorithms has been.

Hasan and Zaki (2011) make an overview of the current techniques used in link prediction and classify them into 3 types of algorithms:

1. The first class of algorithms computes a similarity score between the nodes and employs a training/learning method; these models are considered as having a classic approach.
2. The second class of algorithms is based on Bayesian probabilistic inference and on probabilistic relational methods.
3. The third class of algorithms is based on graph evolution models or linear algebraic formulations.

Besides proposing this taxonomy for link prediction, Hasan and Zaki (2011) also identify the current problems and research gaps in this field. Specifically, they show that time-aware link prediction (or predicting the evolution of a network topology), scalability of proposed solutions (particularly in the case of probabilistic algorithms), and game theoretic approaches to link prediction are areas where more research is needed.

Liben-Nowell and Kleinberg (2007) proposed some of the earliest link prediction methods in a social networks, based on node proximity, and compare them with other algorithms by ranking them on their accuracy and performance. They looked at various similarity measurements between pairs of nodes. Their proposed methodology is part of the first class of algorithms that uses similarity between known nodes in order to train the model for future nodes and compared the performance of algorithms such as Adamic/Adar, weighted Katz, Katz clustering, low-rank approximation (inner product), Jaccard's coefficient, graph distance, common neighbors, hitting time, rooted PageRank, SimRank, and compared these on five networks of co-authorship from

arXiv and concluded that the best performance was given by the Katz clustering, although the Katz, Adamic/Adar, and low-rank inner product are similar in their predictions. They also found that the most different method from all others was the hitting time. Nonetheless, all algorithms perform quite poorly, with only a 16% accuracy as the maximum best prediction from Katz on only one data set.

Sharma et al. (2014) also review the current techniques used in link prediction and make an experimental comparison between these. They also classify these techniques in 3 groups:

1. Node based techniques.
2. Link based techniques.
3. Path based techniques.

They also classify the link prediction techniques as Graph theoretic approach, Statistical approach, Supervised learning approach, and Clustering and they choose 12 of the most used techniques that they classify and test experimentally. These 12 techniques are the following: Node Neighborhood, Jaccard's Coefficient, Adamic/Adar, Hitting Time, Preferential Attachment, Katz ($\alpha = 0.01$), Katz ($\alpha = 0.001$), Katz ($\alpha = 0.0001$), Sim Rank, Commute Time, Normalized Commute Time, LRW, SRW, Rooted Pagerank ($\alpha = 0.01$), Rooted Pagerank ($\alpha = 0.1$), Rooted Pagerank ($\alpha = 0.5$). They compared the precision of the link prediction from 12 techniques on a real dataset and concluded that the Local Random Walk (LRW) technique has the best performance.

Lü and Zhou (2011) also surveyed a series of techniques used in link prediction, from the statistical physicist point of view. They classify the algorithms similarly as Hasan and Zaki (2011), as:

1. Similarity based algorithms
2. Maximum Likelihood Methods
3. Probabilistic models

They also emphasize that most of the current techniques are focused on the unweighted undirected networks and that directed networks add another layer of complexity to the problem. Also,

another difficult problem is to predict not only the existence of a link, but also the weight of that link.

They also show that more challenges in link prediction come from multi-dimensional networks. Specifically, a big challenge is the link prediction in multi-dimensional networks, where links could have different meanings or where the network is consisted of several classes of nodes.

Link prediction becomes a particular problem in the case of sparse networks (Lichtenwalter et al. 2010). The authors address it using a “supervised approach” through training classifiers. They also dismiss the unsupervised approaches, which are based on node-neighborhoods or path information as too simplistic, since they are based on a single metric (Lichtenwalter et al. 2010). On the contrary, the supervised approaches are using training classifiers.

Another research shows that link prediction can be effectively done by using a spatial proximity approach and not network-based measures (Wang et al. 2011).

Particularly in very large datasets or very large complex networks, link prediction is a critical algorithm for understanding the evolution of such networks and their dynamic topology, particularly in social media data, where the links can be sparse or missing and there is an abundance of nodes and information exchanged through these nodes.

Additionally, link prediction algorithms have been more recently used also to improve the performance of graph neural networks (Zhang and Chen 2018) and show great potential for refinement or current neural networks and AI algorithms.

Further Reading

- Al Hasan, M., & Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics* (pp. 243–275). Boston: Springer.
- Liben-Nowell, D., & Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), 1019–1031.
- Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD*

international conference on Knowledge discovery and data mining. ACM.

- Linyuan Lü, & Zhoua, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6), 1150–1170.
- Sharma, D., Sharma, U., & Khatri, S. K. (2014). An experimental comparison of the link prediction techniques in social networks. *International Journal of Modeling and Optimization*, 4(1), 21.
- Wang, D., et al. (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.
- Wei, X., Xu, L., Cao, B., & Yu, P. S. (2017). Cross view link prediction by learning noise-resilient representation consensus. In *Proceedings of the 26th International Conference on World Wide Web*, 1611–1619.
- Zhang, M., & Chen, Y. (2018). Link prediction based on graph neural networks. In *Advances in neural information processing systems*.
- Zhou, M.-Y., Liao, H., Xiong, W.-M., Wu, X.-Y., & Wei, Z.-W. (2017). Connecting Patterns Inspire Link Prediction in Complex Networks. *Complexity*, 2017., Article ID 8581365, 12 p. <https://doi.org/10.1155/2017/8581365>.

Link/Graph Mining

Derek Doran

Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA

Synonyms

Network analysis; Network science; Relational data analytics

Definition/Introduction

Link/graph mining is defined as the extraction of information within a collection of interrelated objects. Whereas conventional data mining imagines a database as a collection of “flat” tables, where entities are rows and attributes of these entities are columns, link/graph mining imagines entities as nodes or vertices in a network, with attributes attached to the nodes themselves.

Relationships among datums in a “flat” database may be seen by primary key relationships or by common values across a set of attributes. In the link/graph mining view of a database, these relationships are made explicit by defining links or edges between vertices. The edges may be *homogeneous*, where a single kind of relationship defines the edges that are formed, or *heterogeneous*, where multiple kinds of data are used to develop a vertex set, and relationships define edges among network vertices. For example, a relation from vertex A to B and a relation from vertex C to D in a homogeneous graph means that A is related to B in the same way that C is related to D. An example of a homogeneous graph may be one where nodes represent individuals and connections represent a friendship relationship. An example of a heterogeneous graph is one where different types of network devices connect to each other to form a corporate intranet. Different node types correspond to different device types, and different relationships may correspond to the type of network protocol that two devices use to communicate with each other. Networks may be **directed** (e.g., a link may be presented from A to B but not vice versa) or **undirected** (e.g., a link from A to B exists if and only if a link from B to A exists). Link/graph mining is intimately related to **network science**, which is the scientific study of the structure of complex systems. Common link/graph mining tasks include discovering shortest or expected paths in the network, an importance ranking of nodes or vertices, understanding relationship patterns, identifying common clusters or regions of a graph, and modeling propagation phenomena across the graph. Random graph models give researchers a way to identify whether a structural or interaction pattern seen within a dataset is statistically significant.

Network Representations of Data

While a traditional “tabular” representation of a dataset contains information necessary to understand a big dataset, a network representation makes explicit datum relations that may be

implicit in a data table. For example, in a database of employee personal and their meeting calendars, a network view may be constructed where employees are nodes and edges are present if two employees will participate in the same meeting. The network thus captures a “who works with who” relationship that is only implicit in the data table. Analytics over the network representation itself can answer queries such as “how did somebody at meeting C hear about information that was only discussed during meeting A?”, or “which employee may have been exposed to the most amount of potential information, rumors, and views, as measured by participating in many meetings where few other participants overlap?” The network representation of data has another important advantage: the network itself represents the *structure* of a complex system of interconnected participants. These participants could be people or even components of a physical system. There is some agreement in the scientific community that the complexity of most technological, social, biological, and natural systems is best captured by its representation as a network.

The field of **network science** is devoted to the scientific application of link and graph mining techniques to quantitatively understand, model, and make predictions over complex systems. Network science defines two kinds of frameworks under which link/graph mining is performed: (i) exploratory analysis and (ii) hypothesis-driven analysis. In exploratory analysis, an analyst has no specific notion about why and how nodes in a complex system connect or are related to each other or why a complex network takes on a specific structure. Exploratory analysis leads to a hypothesis about an underlying mechanism of the system based on regularly occurring patterns or based on anomalous graph metrics. In hypothesis-driven analysis, the analyst has some at hand evidence supporting an underlying mechanism about how a system operates and is interested in understanding how the structural qualities of the system speak in favor or in opposition to the mechanism. Under either setting, hypotheses may be tested by comparing observations against random network models to identify whether or not patterns in support or in opposition of a

hypothesis are significant or merely occurred by chance. Network science is intimately tied to link/graph mining: it defines an apparatus for analysts to use link/graph mining methods that can answer important questions about a complex system. Similarly, network science procedures and analyses are the primary purpose for the development of link/graph mining techniques. The utility of one would thus not nearly be as high without the other.

Representation

The mathematical representation of a graph is a basic preprocessing step for any link/graph mining task. One form may be as follows: every node in the graph is labeled with an integer $i = 1 \dots n$ and a tuple (i, j) is defined for a relationship between nodes i and j . A network may then be defined by the value n and a list of all tuples. For example, let $n = 5$ and define the set $\{(1, 2), (3, 4), (2, 4), (4, 1), (2, 3)\}$. This specifies a graph with five vertices, one of which is disconnected (vertex 5) and the others that have edges between them as defined by the set. Such a specification of a network is called an **edge list**. Another approach is to translate the edge list representation into an **adjacency matrix** $\mathbf{A}$. This is defined as an $n \times n$ matrix where the element A_{ij} , corresponding to the i^{th} row and j^{th} column of the matrix, is equal to 1 if the tuple (i, j) or (j, i) exists in the edge list. When edges are unlabeled or unweighted, $\mathbf{A}$ is simply a binary matrix. Alternatively, if the graph is heterogeneous or allows multiple relationships between the same pair of nodes, then A_{ij} is equal to the number of edges between i and j . When $\mathbf{A}$ is not symmetric, the graph is directed rather than **undirected**.

Types of Link/Graph Mining Techniques

The discovery and analysis of algorithms for extracting knowledge from networks are ongoing. Common *types* of analyses, emphasizing those types often used in practice, are explained below.

Path analysis: A **path** p in a graph is a sequence of vertices $p = (v_1, v_2, \dots, v_m)$, $v_i \in V$ such that for each consecutive pair v_i, v_j of vertices in p is matched by an edge of the form (v_j, v_i) (if the network is undirected) or (v_i, v_j) (if the network is

directed or undirected). If one were to draw a graph graphically, a path is any sequence of movements along the edges of the network that brings you from one vertex to another. Any path is valid, even ones that have loops or crosses the same vertex many times. Paths that do not intersect with themselves (i.e., v_i does not equal v_j for any $v_i, v_j \in p$) are self-avoiding. The length of a path is defined by the total number of edges along it. **Geodesic paths** between vertices i and j is a minimum length path of size k where $p_1 = i$ and $p_k = j$. A **breadth-first search** starting from node d , which iterates over all paths of length 1, and then 2 and 3, and so on up to the largest path that originates at d , is one way to compute geodesic paths.

Network interactions: Whereas path analysis considers the global structure of a graph, the interactions among nodes are a concept related to subgraphs or microstructures. Microstructural measures consider a single node, members of its n^{th} **degree neighborhood** (the set of nodes no more than n hops from it), and the collection of interactions that run between them. If macro-measures study an entire system as a whole (the “forest”), micro-measures such as interactions try to get at the heart of the individual conditions that cause nodes to bind together locally (the “trees”). Three popular features for microstructural analysis are **reciprocity**, **transitivity**, and **balance**.

Reciprocity measures that degree to which two nodes are mutually connected to each other in a directed graph. In other words, if one observes that a node A connects to B , what is the chance that B will also connect A ? The term reciprocity comes from the field of **social network analysis**, which describes a particular set of link/graph mining techniques designed to operate over graphs where nodes represent people and edges represent the social relationships among them. For example, if A does a favor for B , will B also do a favor for A ? If A sends a friend request to B on an online social system, will B reply? On the World Wide Web, if website A has a hyperlink to B , will B link to A ?

Transitivity refers to the degree to which two nodes in a network have a mutual connection in common. In other words, if there is an edge between nodes A and B and B to C , graphs that

are highly transitive indicate a tendency for an edge to also exist between A and C. In the context of social network analysis, transitivity carries an intuitive interpretation based on the old adage “a friend of my friend is also my friend.” Transitivity is an important measure in other contexts, as well. For example, in a graph where edges correspond to paths of energy as in a power grid, highly transitive graphs correspond to more efficient systems compared to less transitive ones: rather than having energy take the path A to B to C, a transitive relation would allow a transmission from A to C directly. The transitivity of a graph is measured by counting the total number of closed triangles in the graph (i.e., counting all subgraphs that are complete graphs of three nodes) multiplied by three and divided by the total number of connected triples in the graph (e.g., all sets of three vertices A, B, and C where at least the edges (A,B) and (B,C) exist).

Balance is defined for networks where edges carry a binary variable that, without loss of generality, is either “positive” (i.e., a “+,” “1,” “Yes,” “True,” etc.) or “negative” (i.e., a “−,” “0,” “No,” “False,” etc.). Vertices incident to positive edges are harmonious or non-conflicting entities in a system, whereas vertices incident to negative edges may be competitive or introduce a tension in the system. Subgraphs over three nodes that are complete are **balanced** or **imbalanced** depending on the assignment of + and − labels to the edges of the triangle as follows:

- Three positive: Balanced. All edges are “positive” and in harmony with each other.
- One positive, two negative: Balanced. In this triangle, two nodes exhibit a harmony, and both are in conflict with the same other. The state of this triangle is “balanced” in the sense that every node is either in harmony or in conflict with all others in kind.
- Two positive, one negative: Imbalanced. In this triangle, node A is harmonious with B, and B is harmonious with C, yet A and C are in conflict. This is an imbalanced disagreement since, if A does not conflict with B, and B does not conflict with C, one would expect A to also not conflict with C. For example, in a social

context where positive means friend and negative means enemy, B can fall into a conflicting situation when friends A and C disagree.

- Three negative: In this triangle, all vertices are in conflict with one another. This is a dangerous scenario in systems of almost any context. For example, in a dataset of nations, mutual disagreements among three states has consequence to the world community. In a dataset of computer network components, three routers that are interconnected but in “conflict” (e.g., a down connection or a disagreement among routing tables) may lead to a system outage.

Datasets drawn from social process always tend toward balanced states because people do not like tension or conflict. It is thus interesting to use link/graph mining to study social systems where balance may actually not hold. If a graph where most triangles are not balanced comes from a social system, one may surmise that there exist latent factors pushing the system toward imbalanced states. A labeled complete graph is balanced if every one of its triangles is balanced.

Quantifying node importance: The **importance** of a node is related to its ability to reach out or connect to other nodes. A node may also be important if it carries a strong degree of “flow,” that is, if the values of relationships connected to it are very high (so that it acts as a strong conduit for the passage of information). Nodes may be important if they are vital to maintain network connectivity, so that if an important node was removed, the graph may suddenly fragment or become disconnected. Importance may be measured recursively: a node is important if it is connected to other nodes that themselves are important. For example, people who work in the United States White House or serve as Senior Aids to the President are powerful people, not necessarily because of their job title but because they have a direct and strong relationship with the Commander in Chief.

Importance is measured by calculating the **centrality** of a node in a graph. Different centrality measures that encode different interpretations of node importance exist and should thus be selected according to the analysis at hand. **Degree centrality** defines importance as being proportional to the

number of connections a node has. **Closeness centrality** defines importance as having a small average distance to all other nodes in the graph. **Betweenness centrality** defines importance as being part of as many shortest paths in graph from other pairs of nodes as possible. **Eigenvector centrality** defines importance as being connected to not only many other nodes but also to many other nodes that are themselves are important.

Graph partitioning: In the same way that clusters of datums in a dataset correspond to groups of points that are similar, interesting, or signify some other demarcation, vertices in graphs may also be divided into groups that correspond to a common affiliation, property, or connectivity structure. Graph partitioning methods. Graph partitioning takes as an input the number and size of the groups and then searches for the “best” partitioning under these constraints. Community detection algorithms are similar to graph partitioning methods except that they do not require the number and size of groups to be specified a priori. But this is not necessarily a disadvantage to graph partitioning methods; if a graph miner understands the domain from where the graph came from well, or if for her application she requires a partitioning into exactly k groups, graph partitioning methods should be used.

Conclusion

As systems that our society relies on become ever more complex, and as technological advances continue to help us capture the structure of this complexity at high definition, link/graph mining methods will continue to rise in prevalence. As the primary means to understand and extract knowledge from complex systems, link/graph mining methods need to be included in the toolkit of any big data analyst.

Cross-References

- [Computer Science](#)
- [Computational Social Sciences](#)
- [Graph-Theoretic Computations/Graph Databases](#)

- [Mathematics](#)
- [Statistics](#)

Further Reading

- Cook, D. J., & Holder, L. B. (2006). *Mining graph data*. Wiley.
- Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, 7(2), 3–12.
- Lewis, T. G. (2011). *Network science: Theory and applications*. Wiley.
- Newman, M. (2010). *Networks: An introduction*. New York: Oxford University Press.
- Philip, S. Y., Han, J., & Faloutsos, C. (2010). *Link mining: Models, algorithms, and applications*. Berlin: Springer.

LinkedIn

Jennifer J. Summary-Smith

Florida South Western State College, Fort Myers, FL, USA

Culver-Stockton College, Canton, MO, USA

According to its website, LinkedIn is the largest professional network in the world servicing over 300 million members in over 200 territories and countries. Their mission statement is to “connect the world’s professionals to make them more productive and successful. When you join LinkedIn, you get access to people, jobs, news, updates, and insights that help you be great at what you do.” Through its online service, LinkedIn earns around \$473.2 million from premium subscriptions, marketing solutions, and talent solutions. It offers free and premium memberships allowing people to network, obtain knowledge, and locate potential job opportunities. The greatest asset to LinkedIn is its data, making a significant impact in the job industry.

Company Information

Cofounder Reid Hoffman conceptualized the company in his living room in 2002, launching LinkedIn on May 5, 2003. Hoffman, a Stanford graduate, became one of PayPal’s earliest

executives. After PayPal was sold to eBay, he cofounded LinkedIn. The company had one million members by 2004. Today, the company is ran by chief executive, Jeff Weiner, who is also the former CEO of Yahoo! Inc. LinkedIn's headquarters are located in Mountain View, California, with US offices in Chicago, Los Angeles, New York, Omaha, and San Francisco. LinkedIn also has international offices in 21 locations and its online content is available in 23 languages. LinkedIn currently employs 5,400 full-time employees with offices in 27 cities globally. LinkedIn states that professionals are signing up to join the service at the rate of two new members per second with 67% of its membership located outside of the United States. The fastest growing demographic using LinkedIn are students and recent college graduates, accounting for around 39 million users. LinkedIn's corporate talent solutions product lines and its memberships include all executives from the 2013 Fortune 500 companies and 89 Fortune 100 companies. In 2012, its members conducted over 5.7 billion professionally oriented searches, with three million companies utilizing LinkedIn company pages.

As noted on cofounder Reid Hoffman's LinkedIn account, a person's network is how one stays competitive as a professional, keeping up-to-date on one's industry. LinkedIn provides a space where professionals learn about key trends, information, and transformations of their industry. It provides opportunities for people to find jobs, clients, and other business connections.

Relevance of Data

MIT Sloan Management Review contributing editor, Renee Boucher Ferguson, interviewed LinkedIn's director of relevance science, Deepak Agarwal, who states that relevance science at LinkedIn plays the role of improving the relevancy of its products by extracting information from LinkedIn data. In other words, LinkedIn provides recommendations using its data to predict user responses to different items.

To achieve this difficult task, LinkedIn has relevance scientists who provide an

interdisciplinary approach with backgrounds in computer science, economics, information retrieval, machine learning, optimization, software engineering, and statistics. Relevance scientists work to improve the relevancy of LinkedIn's products. According to Deepak Agarwal, LinkedIn relevance scientists significantly enhance products such as advertising, job recommendations, news, LinkedIn feed, people recommendations, and much more. He further points out that most of the company's products are based upon its use of data.

Impact on the Recruiting Industry

As it states on LinkedIn's website, the company's free membership allows its members the opportunity to upload resumes and/or curriculum vitae, join groups, follow companies, establish connections, view and/or search for jobs, endorse connections, and update profiles. It also suggests to its members several people that they may know, based on their connections. LinkedIn's premium service provides members with additional benefits, allowing access to hiring managers and recruiters. Members can send personalized messages to any person on LinkedIn. Additionally, members can also find out who has viewed their profile, detailing how others found them for up to 90 days. There are four premium search filters, permitting premium members to find decision makers at target companies. The membership also provides individuals the opportunity to get noticed by potential employers. When one applies as a featured applicant, it raises his or her rank to the top of the application list. OpenLink is a network that also lets any member on LinkedIn to view another member's full profile to make a connection.

The premium LinkedIn membership assists with drawing attention to members' profile, adding an optional premium or job seeker badge. When viewing the job listings, members have the option to sort by salary range, comparing salary estimates for all jobs in the United States, Australia, Canada, and the United Kingdom. LinkedIn's premium membership also allows

users to see more profile data in one's extended network, including first-, second-, and third-degree connections. A member's first-level connections are people that have either received an invitation from the member or the member sent an invitation to connect. Second-level connections are people who are connected to first-level connections but are not connected to the actual member. Third-level connections are only connected to the second-level members. Moreover, members can receive advice and support from a private group of LinkedIn experts, assisting with job searches.

In a recent article by George Anders, he notes the impact that LinkedIn has made on the recruiting industry. He spoke with the chief executive of LinkedIn, Jeff Weiner, who brushes off comparisons between LinkedIn and Facebook. While both companies connect a vast amount of people via the Internet, each social media platform occupies a different niche within the social networking marketplace. Facebook generates 85% of its revenue from advertisements, whereas LinkedIn focuses its efforts on monetizing members' information. Furthermore, LinkedIn's mobile media experience is growing significantly, changing the face of job searching, career networking, and online profiles. George Anders also interviewed the National Public Radio head of talent acquisition, Lars Schmidt, who notes that recruiters no longer remain chiefly in their offices but are becoming more externally focused. The days of exchanging business cards is quickly being replaced by smartphone applications such as CardMunch. CardMunch is an iPhone app that captures business card photos, transferring them into digital contacts. In 2011, LinkedIn bought the company, retooling it to pull up existing LinkedIn profiles from each card improving the ability of members to make connections. A significant part of LinkedIn's success comes from its dedication to selling services to people who purchase talent.

The chief executive of LinkedIn, Jeff Weiner, has created an intense sales-focused culture. The company celebrates new account wins during its biweekly meetings. According to George Anders, LinkedIn has doubled the number of sales employees in the past year. In addition, the

company has made a \$27 billion impact on the recruiting industry. Jeff Weiner also states that every time LinkedIn expands its sales team for hiring solutions, the payoff increases "off the charts." He also talks about how sales keep rising and its customers are spreading enthusiasm for LinkedIn's products. Jeff Weiner further states that once sales are made, LinkedIn customers are loyal, reoccurring, and low maintenance. This trend is reflected in current stock market prices in the job-hunting sector. George Anders writes that older search firm companies, such as Heidrick & Struggles that recruits candidates the old fashion way, have slumped 67%. Monster Worldwide has experienced a more dramatic drop, tumbling 81%.

As noted on its website, "LinkedIn operates the world's largest professional network on the Internet." This company has made billions of dollars, hosting a massive amount of data with a membership of 300 million people worldwide. The social network for professionals is growing at a fast pace under the tenure of Chief Executive Jeff Weiner. In a July 2014 article by David Gelles, he reports that LinkedIn has made its second acquisition in the last several weeks buying Bizo for \$175 million dollars. A week prior, it purchased Newsle, which is a service that combs the web for articles that are relevant to members. It quickly notifies a person whenever friends, family members, coworkers, and so forth are mentioned online in the news, blogs, and/or articles.

LinkedIn continues to make great strides by leveraging its large data archives, to carve out a niche in the social media sector specifically targeting the needs of online professionals. It is evident that, through the use of big data, LinkedIn is changing and significantly influencing the job-hunting process. This company provides a service that allows its member to connect and network with professionals. LinkedIn is the world's largest professional network, proving to be an innovator in the employment service industry.

Cross-References

- [Facebook](#)
- [Information Society](#)

- [Online Identity](#)
- [Social Media](#)

Further Reading

Anders, G. *How LinkedIn has turned your resume into a cash machine*. <http://www.forbes.com/sites/georgeanders/2012/06/27/how-linkedin-strategy/>. Accessed July 2014.

Boucher Ferguson, R. *The relevance of data: Behind the scenes at LinkedIn*. <http://sloanreview.mit.edu/article/the-relevance-of-data-going-behind-the-scenes-at-linkedin/>. Accessed July 2014.

Gelles, D. *LinkedIn makes another deal, buying Bizo*. http://dealbook.nytimes.com/2014/07/22/linkedin-does-another-deal-buying-bizo/?_php=true&_type=blogs&_php=true&_type=blogs&_php=true&_type=blogs&r=2. Accessed July 2014.

LinkedIn. <https://www.linkedin.com>. Accessed July 2014.

Machine Intelligence

► Artificial Intelligence

Machine Learning

Ashrf Althbiti and Xiaogang Ma
Department of Computer Science, University of
Idaho, Moscow, ID, USA

Introduction

Machine learning (ML) is a fast-evolving scientific field that effectively copes with big data explosion and forms a core infrastructure for artificial intelligence and data science. ML bridges the research fields of computer science and statistics and builds computational algorithms and statistical model-based theories from those fields of studies. These algorithms and models are utilized by automated systems and computer applications to perform specific tasks, with the desire of high prediction performance and generalization capabilities (Jordan and Mitchell 2015). Sometimes, ML is also referred to as a predictive analytics or statistical learning. The general workflow of a ML system is that it receives inputs (aka, training sets), trains predictive models, performs specific prediction tasks, and eventually generates outputs. Then, the ML

system evaluates the performance of predictive models and optimizes the model parameters in order to obtain better predictions. In practice, ML systems also learn from prior experiences and generate solutions for given problems with specific requirements.

Machine Learning Approaches

Jordan and Mitchell (2015) discussed that the main paradigms of ML methods are (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning. ML approaches are categorized based on two criteria: (1) the data type of a dependent variable and (2) the availability of labels of a dependent variable. The former criterion is categorized into two classes: (1) continue and (2) discrete. The latter criterion is utilized to determine the type of ML algorithm. If a dependent variable is given and labeled, it would be a supervised learning approach. Otherwise, if a dependent variable is not given or unlabeled, it would be an unsupervised learning approach.

Supervised learning algorithms are often utilized for prediction tasks and building a mathematical model of a set of data that include both inputs and desired outputs. These algorithms learn a prediction model that approximates a function $f(x)$ to predict an output y (Hastie et al. 2009). For instance, fraud classifier of credit-card transactions, spam classifier of emails, and medical diagnosis systems (e.g., breast cancer diagnosis) each

represents a function of approximation that supervised learning algorithms perform.

Unsupervised learning algorithms are often used to study and analyze a dataset and learn a model that finds and discovers a useful structure of the inputs without the need of labeled outputs. They are also used to address two major problems that researchers encounter in the ML workflow: (1) data sparsity where missing values can affect a model's accuracy and performance and (2) curse of dimensionality which means data is organized in high-dimensional spaces (e.g., thousands of dimensions).

Reinforcement learning forms a major ML paradigm where it sits at the crossroad of supervised and unsupervised learning (Mnih et al. 2015). For reinforcement learning, the availability of information in training examples is intermediate between supervised and unsupervised learning. In another words, the training examples provide indications about an output inferred by the correctness of an action. Yet, if an action is not correct, the challenge of finding a correct action endures (Jordan and Mitchell 2015).

Other ML approaches emerge when researchers develop combinations across the three main paradigms, such as semi-supervised learning, discriminative training, active learning, and causal modeling (Jordan and Mitchell 2015).

Machine Learning Models

As a result of utilizing ML algorithms on a training dataset, a model is learned to make predictions on new datasets. Table 1 lists different ML algorithms based on the availability of labeled output variables and their data types. Table 2 gives a longer list of the state-of-the-art models in ML algorithms (Amatriain et al. 2011).

Classification and Regression

The following algorithms are briefly introduced.

Linear Regression

Linear regression is a statistical model for modeling the relationship between a dependent variable

Machine Learning, Table 1 Classification of machine learning models

	Data are labeled (supervised learning)	Data are unlabeled (unsupervised learning)
Continuous	Regression	Dimensionality reduction
Discrete	Classification	Clustering

y and one or more independent variables X . Formula (1) is a linear model for several explanatory variables represented by a hyperplane in higher dimensions:

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \dots + w[p] * x[p] + b \quad (1)$$

where $x[0]$ to $x[p]$ signify features of a single instance and w and b are learned parameters by minimizing the *mean squared error* between predicted values $\hat{y}$ and true values of y on the training set. Linear regression also forms other models such as ridge and LASSO. Moreover, a regression model, namely, logistic regression, can be applied for classification where target values are transformed into two classes as the prediction formula (2) shows:

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \dots + w[p] * x[p] + b > 0 \quad (2)$$

where the threshold of a predicted value is zero. Thus, if a predicted value is greater than zero, a predicted class is +1; otherwise it is -1.

K-Nearest Neighbors (KNNs)

K-nearest neighbor models are utilized to make a prediction for a new single point (aka, instance). It is utilized for classification and regression problems and is known as lazy learner because it needs to memorize the training sets to make a new prediction (aka, instance-based learning). This model makes a prediction for a new instance based on the values of the nearest neighbors. It finds those nearest neighbors by calculating similarities and distances between a

Machine Learning, Table 2 Different categories of ML models

ML paradigm	Task	Type of algorithm	Model
Supervised	Prediction	Regression and classification	Linear regression
			Ridge regression
			Least absolute shrinkage and selection operator (LASSO)
			<i>K</i> -nearest neighbors for regression
			<i>K</i> -nearest neighbors for classification
			(logistic regression)
			One vs. rest linear model for multi-label classification
			Decision trees (DSs)
			Bayesian classifiers
			Support vector machines (SVMs)
Unsupervised	Features extraction	Dimensionality reduction	Artificial neural networks (ANNs)
			Principal component analysis (PCA)
	Description	Clustering	Singular value decomposition (SVD)
			<i>k</i> -means
			Density-based spatial clustering of application with noise (DBSCAN)
			Message passing
			Hierarchical
		Association rule mining	A priori

single point and its neighbors. The similarity is calculated using Pearson correlation, cosine similarity, Euclidian distance, or other similarity measures.

Decision Trees (DSs)

Decision trees classify a target variable in the form of a tree structure. The nodes of a tree can be (1) decision nodes, where their values are tested to determine to which branch a subtree moves, or (2) leaf nodes, where a class of a data point is determined. Decision nodes must be carefully selected to enhance the accuracy of prediction. DSs can be used in regression and classification applications.

Bayesian Classifiers

Bayesian classifiers are a probabilistic framework to address classification and regression needs. It is based on performing Bayes’ theorem and the definition of conditional probability. The main assumption of applying Bayes’ theorem is that features should maintain strong (naïve) independence.

Support Vector Machines (SVMs)

Support vector machines are classifiers that strive to separate data points by finding linear hyper-planes that maximize margins between data points in an input space. It is noteworthy that SVMs can be applied to address regression and classification needs. The support vectors are data points that fit on maximized margins.

Artificial Neural Networks (ANNs)

Artificial neural networks are models that inferred from biological neural networks of the brain. It develops a network of interconnected neurons which work together to perform prediction tasks. Numerical weights are assigned to the links between nodes and are tuned based on experience. The simple representation network consists of three main layers: (1) input layer, (2) hidden layer, and (3) output layer. Handwriting recognition is a typical application where ANNs are used.

Dimensionality Reduction

The poor performance of ML algorithms is often caused by the number of dimensions in a data

space. Hence, an optimal solution is to reduce the number of dimensions, while the maximum amount of information is retained. PCA and SVD are the main ML algorithms that offer a solution to the issue of dimensionality.

Clustering

Clustering is a popular ML algorithm which falls in the unsupervised learning category. It groups data points based on their similarity. Thus, data points that fit in one cluster or class are different from the data points in another cluster. A common technique of clustering is a k -means where k indicates total number of clusters. The k -means clustering algorithm randomly selects k -number of data points and plots them on a Cartesian plane. These data points form a centroid of each cluster where the remaining data points are assigned to the best centroid. Then, a process of reassigning centroids is repeated inside each cluster until there are no more changes in a set of k centroids. Other ML algorithms considered as an alternative selection of k -means are DBSCAN, message-passing clustering, hierarchical clustering, etc.

Association Rule Mining

Association rule mining algorithms are mostly used in marketing when predicting co-occurrence of items in a transaction. It is widely utilized to identify co-occurrence relationship patterns in large-scale data points (e.g., items or products).

Model Fitness and Evaluation

The main objective of adopting ML algorithms on training dataset is to generalize a learned model to make accurate predictions on new data points. Hence, if a model makes accurate predictions on new data points, this model would be generalized from a training dataset to test datasets. However, an extensive training of a model increases the complexity, in which the overfitting problem may appear. The overfitting problem means that a model does memorize the training dataset and perform well on training dataset, but is not able to make accurate

predictions on test datasets. On the other hand, if a model is not sufficiently trained on a training dataset, this model most likely will do badly even on a training dataset. Hence, the goal is to select a model that maintains an optimal complexity of training.

Learning a model requires a set of data points as inputs to train a model, a set of data points to tune and optimize a model's parameters, and a set of data points to evaluate its performance. Therefore, a dataset is divided into three sets, namely, the training set, the evaluating set, and the testing set. The way of dividing these sets depends on algorithm developers. There are different techniques to be followed when dividing datasets. One basic technique is to utilize a 90/10 rule of thumb which means 90% of a dataset is used to learn a model and the other 10% is used to evaluate and adjust it. Other methods for dataset splitting include k -fold cross-validation and hold-out cross-validation (Picard and Berk 2010). Furthermore, there are other sophisticated statistical evaluation techniques applicable for different types of datasets, such as bootstrapping methods which depend on random sampling with replacement or grid search.

A wide range of evaluation criteria can be used for evaluating ML algorithms. For example, accuracy is an extensively utilized property to gauge the performance of model predictions. Typical examples of accuracy measurements are R^2 and root-mean-square error (RMSE). Other metrics to measure the accuracy of usage prediction include precision, recall, support, F-score, etc.

Applications

Various applications have utilized ML techniques to automate inferences and decision-making under uncertainty, which include, but not limited to, health care, education, aerospace, neuroscience, genetics and genomics, cloud computing, business, e-commerce, finance, and supply chain. Within artificial intelligence (AI), the dramatic emergence of ML has been utilized as the method of choice for building systems for computer vision, speech recognition, facial

recognition, natural language processing, and other applications (Jordan and Mitchell 2015).

Conclusion

ML provides models that learn automatically through experience. The explosion of big data is a main motivation behind the evolution of ML approaches. A survey of the current state of ML algorithms is introduced in a coherent fashion in this entry to simplify the rich and detailed content. Also, the discussion has extended to cover more topics that demonstrate how utilizing machine learning algorithms can alleviate the issue of dimensionality and offer solutions to automate prediction and detection. Also, model evaluation and machine learning applications are briefly introduced.

Cross-References

► [Financial Data and Trend Prediction](#)

Further Reading

- Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender systems handbook* (pp. 39–71). Boston: Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*. <https://doi.org/10.1126/science.aaa8415>.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529.
- Picard, R. R., & Berk, K. N. (2010). Data splitting. *American Statistician*, 44(2), 140–147. <https://doi.org/10.1080/00031305.1990.10475704>.

Maritime Data

► [Maritime Transport](#)

Maritime Shipping

► [Maritime Transport](#)

Maritime Transport

Camilla B. Bosanquet

Schar School of Policy and Government, George Mason University, Arlington, VA, USA

Synonyms

[Maritime Data](#); [Maritime Shipping](#)

Ancient Origins of Maritime Transport Data

Maritime transport data collection and dissemination have their roots in antiquity. Mercantile enterprise first necessitated such information, the imperative for which deepened over time with the emergence of passenger movement by sea, coastal defense, and naval operations. Early data took a variety of forms, e. g., Ancient Egyptian tombs and papyri recorded maritime commerce activities as early as 1500 BCE; Zenon of Kaunos, private secretary to the finance minister of Ptolemy II, documented vessel cargo manifests in 250 BCE; and “sailing directions” dating to the third century BCE provided ship captains of the Roman Empire with critical weather, ship routing, and harbor guidance for Indian coastal markets (Casson 1954).

While modern maritime shipping has benefited from remarkable technological advancements during the intervening millennia, the business of maritime transport relies upon timeless concepts, even as the capture and conveyance of relevant data has jumped from paper to digital form. While global positioning satellites replaced the nautical sextant and compass, electronic logbooks supplanted handwritten voyage diaries, and touch-screen navigation displaced

plotting routes and position fixes in pencil on paper charts, surface vessel operation and management still bear a strong resemblance to activities of yesteryear. It is similarly the case with the commercial movement of goods and people between ports; passenger and cargo manifests serve the same essential needs as those of past ages.

Maritime Transport and Big Data

What is new, in modernity, is the immeasurable *quantity* of data related to maritime transport and the lightning *speed* at which such data can proliferate and travel. Ship owners, fleet managers, cargo shipping agents and firms, vessel operators and engineers, government officials, business logisticians, port management, economists, market traders, commodity and energy brokers, financial investors, insurance firms, maritime organizations, vessel traffic service watch standers, classification societies, admiralty lawyers, navies, coast guards, and others all benefit from access to maritime information.

Maritime data can, inter alia, be used by the aforementioned actors in myriad endeavors to:

- Ascertain and predict sea and meteorological conditions
- Obtain the global locations of vessels in real-time
- Make geospatial transit calculations based upon port locations
- Optimize ship routing given weather, port options, available berthing, fuel costs, etc.
- Monitor the status of shipboard engineering plant machinery
- Translate hydrographic survey data to under-sea cartographic charts
- Track shipments of containerized, bulk, break-bulk, liquid, and roll-on/roll-off cargoes
- Document endangered whale and sea turtle sightings for ship strike avoidance
- Calculate search areas and human survival likelihood given currents, winds, temperatures
- Record transit temperatures of refrigerated perishables (e.g., food, flowers, medicine)
- Estimate tsunami wave probabilities resulting from seismic activity
- Enable law enforcement to intercept actors engaged in illicit activity at sea
- Enforce exclusive economic zones, marine sanctuaries, and closed fishing grounds
- Communicate relocated harbor buoy and navigation hazard information promptly
- Contribute to vessels' timely inspection, certification, maintenance, and overhaul
- Study economic flows of goods and people by sea
- Facilitate coastal incident response (e.g., oil spill mitigation, plane crash recovery)
- Develop strategies to avoid, defend against, counter, and withstand maritime piracy
- Evaluate the performance of vessels, equipment, captains, crews, ports, etc.
- Identify vessels and operators violating cabotage laws or embargoes
- Monitor adherence to flag state and port state control requirements
- Determine financial and insurance risks associated with ships and vessel fleets
- Inform strategies for coastal defense, military sealift, and naval projection of power

As of 2005, the United Nations' International Maritime Organization mandated the worldwide operation of automatic identifications systems (AIS) onboard all passenger vessels and certain ships specified by tonnage and/or purpose. Accomplished under the International Convention for the Safety of Life at Sea, the new regulation required the use of each vessel's AIS transponder to transmit ship identification, vessel type, and navigational information to other ships, shore-based receivers, and aircraft, and receive such data from other ships. Geosynchronous satellites can now collect AIS transponder data, enabling the near-instantaneous monitoring of

ships that have exceeded terrestrial receivers' tracking ranges.

While a plethora of global maritime intelligence providers boast AIS-informed vessel tracking and associated analytics, Lloyd's of London arguably offers the most comprehensive maritime data, having been collected from the greatest number of sources. Beyond providing vessel tracking, the Lloyd's List Intelligence *Seasearcher* online user interface provides clients with searchable bills of lading, individual container tracking, vessel sanctions information, registered owner details, shipping companies' credit ratings, identification of dry bulk vessels by products carried, liquid natural gas and oil tanker movements and port calls, multilevel vessel and fleet ownership structures, detailed port characteristics, vessel risk detection, assessment of suspicious activity (e.g., breaking sanctions, fishing illegally, illicit trafficking) during AIS gap periods, incident notifications (e.g., cargo seizures, ship casualties, crew arrests, vessel detentions), and more. At the macro-level, such data and analytics enable clients to forecast market trends, manage risk, evaluate global trade, develop strategy, design unique algorithmic applications, and conduct an unlimited variety of tailored analyses.

The Future of Maritime Transport Data

Increased port and vessel automation, alongside further development of autonomous vessel technologies, should enhance the quality of maritime transport data. Prolific use of sensors throughout the maritime transport industry will strengthen data on weather, vessels, ports, cargoes, human operators, and the performance of automated machines. AI and machine learning applications should improve terrestrial and satellite imagery analyses, providing stakeholders with a greater understanding of vessels' activities and vulnerabilities. More timely data should facilitate financial gains or savings as executives and vessel captains make informed decisions concerning efficient ship routing, maintenance availabilities,

marine bunker fueling, anti-piracy voyage planning, ship charter pricing, and freight fees. Aforementioned public-sector entities will likewise experience benefits, e. g., a reduction in marine casualties, fewer marine environmental incidents, and greater intelligence concerning transnational organized crime activity in the maritime domain.

A feasible scenario emerges from this vision: a virtuous feedback loop in which improved maritime technologies refine maritime transport data and multiply end-user benefits, leading to greater investment in maritime technologies, further refining data and boosting benefits. In many respects, big data analytics of maritime transport information is the latest manifestation of an urge that once prompted the creation of "sailing directions" to India during the days of the Roman Empire. It is a human desire to collect, analyze, and disseminate critical information to facilitate the safe and profitable transport of people and goods over the sea.

Cross-References

- [Business Intelligence](#)
- [Intelligent Transportation Systems \(ITS\)](#)
- [Spatiotemporal Analytics](#)

Further Reading

- Casson, L. (1954). Trade in the ancient world. *Scientific American*, 191(5), 98–104.
- Fruth, M., & Teuteberg, F. (2017). Digitization in maritime logistics – What is there and what is missing? *Cogent Business & Management*, 4. <https://doi.org/10.1080/23311975.2017.1411066>.
- Jović, M., Tijan, E., Marx, R., & Gebhard, B. (2019). Big data management in maritime transport. *Pomorski Zbornik*, 57(1), 123–141. <https://doi.org/10.18048/2019.57.09>.
- Notteboom, T., & Haralambides, H. (2020). Port management and governance in a post-COVID-19 era: Quo vadis? *Maritime Economics & Logistics*, 22(3), 329–352. <https://doi.org/10.1057/s41278-020-00162-7>.
- Stopford, M. (2009). *Maritime economics*. London: Routledge.

Mathematics

Daniele C. Struppa

Donald Bren Presidential Chair in Mathematics,
Chapman University, Orange, CA, USA

Introduction

The term *big data* refers to data sets that are so large and complex as to make traditional data management insufficient. Because of the rapid increase in data gathering, in data storage, as well as in the techniques used to manage data, the notion of big data is somewhat dependent on the capability and technology of the user. Traditionally, the challenge of big data has been described in terms of the 3Vs: high Volume (the size and amount of data to be managed), high Velocity (the speed at which data need to be acquired and handled), and high Variety (the range of data types to be managed). Specifically, “Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value (De Mauro et al. 2016).” More recently, some authors have added two more dimensions to the notion of big data, by introducing high Variability (inconsistency in the data set) and high Veracity (the quality and truthfulness of data present high variability, thus making its management more delicate).

The Growth of Data

Big data (its collection, management, and manipulation) has become a topic of significant interest because of the rapid growth of available data due to the ease of access to and collection of data. While until a couple decades ago big data used to appear mostly through scientific data collection, we are now at a point in which each and every digital interaction (through mobile devices,

home computers, cameras in public areas, etc.) generates a digital footprint, which quickly creates extremely large amount of data. One could say that information and digitalization are the fuels that allow data to grow into big data. If the storage of such data is an immediate challenge, even more challenging is to find efficient ways to process such data and finally ways to transmit them. Despite these challenges, big data are now considered a fundamental instrument in a variety of scientific fields (e.g., to forecast the path of weather events such as hurricanes, to model climate changes, to simulate the behavior of blood flow when new medical devices such as stents are inserted in arteries, or finally to make predictions on the outcome of a disease through the use of genomics and proteomics assays), as well as in many business applications (data analytics is currently one of the areas of fastest growth, because of its implied success in determining customer preferences, as well as the risk of financial decisions).

The Power of Data and Their Impact on Science and Mathematics

A few authors (Weinberger 2011) have argued that it will be increasingly possible to answer any question of interest, as long as we have enough raw data about the question being posed. This philosophical assumption has been called the *microarray paradigm* in Napolitano et al. (2014), and it has important implications for the way in which both mathematics and science develop and interact. While in the past scientific theories were seen as offering a theoretical model that would describe reality (think, e.g., of Newtonian mechanics, which explains the behavior of bodies on the basis of three relatively simple laws), the advent of the use of big data seems to herald an era of *agnostic science* (Napolitano et al. 2014), in which mathematical techniques are used to allow the scientist (or the social scientist) to make predictions as to the outcome of a process, even in the absence of a model that

would explain the behavior of the system at hand. The consequence of this viewpoint is the development of new techniques, originating in the field one could call computational mathematics, whose validity is demonstrated not through the traditional methods of demonstrations and proofs but rather by its own applicability to a given problem.

Mathematical Techniques

Mathematicians have employed a wide array of mathematical techniques to work with large data sets and to use them to make inferences.

One of the most successful theories that allow the utilization of large data sets in a variety of different disciplines in the spirit of the agnostic science we referred to goes under the name of statistical learning theory. With this terminology, we mean that particular approach to machine learning that takes its tools from the fields of statistics and functional analysis. In particular, statistical learning theory embraces the approach of supervised learning, namely, learning from a training set of data.

Supervised learning can really be seen as an interpolation problem. Imagine every point in the data set to be an input-output pair (e.g., a genomic structure and a disease): the process of learning consists then in finding a function that interpolates these data and then using it to make predictions when new data are added to the system. While the theory of statistical learning is very developed, among the specific techniques that are used, we will mention clustering techniques such as affinity propagation method, as described, for example, in Frey and Duek (2007). In this case the idea is to split the data into clusters, but passing information locally among various data points, in order to determine the split. Once that is done, the method extracts a best representative from each cluster. The interpolation process is then used on those particular representatives.

Another classical example of mathematical technique that is employed to study large data

sets goes under the name of boosting (Schapire 1990), where a large number of mediocre (slightly better than random) classifiers are combined to provide a much more robust classifier.

Another relatively new addition to the toolkit of the practitioner of big data management is what goes under the name of nonlinear manifold learning (Roweis and Saul 2000), intended as an entire set of techniques designed to find low-dimensional objects that preserve some important structural properties in an otherwise collection of high-dimensional data.

Finally we will mention the growing use of neural networks and of techniques that use the fundamental ideas behind such tool. The idea that originally inspires neural networks consisted in designing networks that would somehow mimic the behavior of biological neural networks, in order to either solve or approximate the solution to the interpolation problem we described above. As the models of neural networks developed, however, the connection with biological networks remained more of an inspiration than an actual guide, and modern neural networks are designed without any such reference. Both supervised and unsupervised learning can occur with the use of neural networks.

Further Reading

- De Mauro, M., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), 122–135.
- Frey, B. J., & Duek, D. (2007). Clustering by passing messages between data points. *Science*, 315, 972–976.
- Napoletani, D., Panza, M., & Struppa, D. C. (2014). Is big data enough? A reflection on the changing role of mathematics in applications. *Notices of the American Mathematical Society*, 61(5), 485–490.
- Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2323–2326.
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2), 197–227.
- Weinberger, D. (2011). The machine that would predict the future. *Scientific American*, 305, 52–57.

Media

Colin Porlezza

IPMZ - Institute of Mass Communication and Media Research, University of Zurich, Zürich, Switzerland

Synonyms

Computer-assisted reporting; Data journalism; Media ethics

Definition/Introduction

Big data can be understood as “the capacity to search, aggregate and cross-reference large data sets” (Boyd and Crawford 2012, p. 663). The proliferation of large amounts of data concerns the media in at least three different ways. First, large-scale data collections are becoming an important resource for journalism. As a result, practices such as data journalism are increasingly gaining attention among newsrooms and become relevant resources as data collected and published in the Internet expands and legal frameworks to access public data such as Freedom of Information Acts come into effect. Recent success stories of data journalism such as uncovering the MPs’ expenses scandal in the UK or the giant data leak in the case of the Panama Papers have contributed to further improve the capacities to deal with large amounts of data in newsrooms. Second, big data are not only important in reference to the practice of reporting. They also play a decisive role with regard to what kind of content gets finally published. Many newsrooms are no longer using the judgment of human editors alone to decide what content ends up on their websites; instead they use real-time data analytics generated by the clicks of their users to identify trends, to see how content is performing, and to boost virality and user engagement. Data is also used in order to improve product development in entertainment formats. Social media like Facebook have perfected this technique by using personal

preferences, tastes, and moods of their users, to offer personalized content and targeted advertising. This datafication means that social media transforms intangible elements such as relationships and transform them into a valuable resource or an economic asset on which to build entire business models. Third, datafication and the use of large amounts of data give also rise to risks with regard to ethics, privacy, transparency, and surveillance. Big data can have huge benefits because it allows organizations to personalize and target products and services. But at the same time, it requires clear and transparent information handling governance and data protection. Handling big data increases the risk of paralyzing privacy, because (social) media or internet-based services require a lot of personal information in order to use them. Moreover, analyzing big data entails higher risks to incur in errors, for instance, when it comes to statistical calculations or visualizations of big data.

Big Data in the Media Context

Within media, big data mainly refers to huge amounts of structured (e.g., sales, clicks) or unstructured (e.g., videos, posts, or tweets) data generated, collected, and aggregated by private business activities, governments, public administrations, or online-based organizations such as social media. In addition, the term big data usually includes references to the analysis of huge bulks of data, too. These large-scale data collections are difficult to analyze using traditional software or database techniques and request new methods in order to identify patterns in such a massive and often incomprehensible amount of data. The media ecosystem has therefore developed specialized practices and tools not only to generate big data but also to analyze it in turn. One of these practices to analyze data is called data or data-driven journalism.

Data Journalism

We live in an age of information abundance. One of the biggest challenges for the media industry, and journalism in particular, is to bring order in

this data deluge. It is therefore not surprising that the relationship between big data and journalism is becoming stronger, especially because large amounts of data need new and better tools that are able to provide specific context, to explain the data in a clear way, and to verify the information it contains. Data journalism is thus not entirely different from more classic forms of journalism. However, what makes it somehow special are the new opportunities given by the combination of traditional journalistic skills like research and innovative forms of investigation thanks to the use of key information sets, key data and new processing, analytics, and visualization software that allows journalists to peer through the massive amounts of data available in a digital environment and to show it in a clear and simple way to the publics. The importance of data journalism is given by its ability to gather, interrogate, visualize, and mash up data with different sources or services, and it requires an amalgamation of a journalist's "nose for news" and tech savvy competences.

However, data journalism is not as new as it seems to be. Ever since organizations and public administrations collected information or built up archives, journalism has been dealing with large amounts of data. As long as journalism has been practiced, journalists were keen to collect data and to report them accurately. When the data displaying techniques got better in the late eighteenth century, newspapers started to use this know-how to present information in a more sophisticated way. The first example of data journalism can be traced back to 1821 and involved *The Guardian*, at the time based in Manchester, UK. The newspaper published a leaked table listing the number of students and the costs for each school in the British city. For the first time, it was publicly shown that the number of students receiving free education was higher than what was expected in the population. Another example of early data journalism dates back to 1858, when Florence Nightingale, the social reformer and founder of modern nursing, published a report to the British Parliament about the deaths of soldiers. In her report she revealed with the help of visual graphics that the main cause of mortality resulted

from preventable diseases during cure rather than as a cause from battles.

By the middle of the twentieth century, newsrooms started to use systematically computers to collect and analyze data in order to find and enrich news stories. In the 1950s this procedure was called computer-assisted reporting (CAR) and is perhaps the evolutionary ancestor of what we call data journalism today. Computer-assisted reporting was, for instance, used by the television network CBS in 1952 to predict the outcome of the US presidential election. CBS used a then famous Universal Automatic Computer (UNIVAC) and programmed it with statistical models based on voting behavior from earlier elections. With just 5% of votes in, the computer correctly predicted the landslide win of former World War II general Dwight D. Eisenhower with a margin of error less than 1%. After this remarkable success of computer-assisted reporting at CBS, other networks started to use computers in their newsrooms as well, particularly for voting prediction. Not one election has since passed without a computer-assisted prediction. However, computers were slowly introduced in newsrooms, and only in the late 1960s, they started to be regularly used in the news production as well.

In 1967, a journalism professor from the University of North Carolina, Philip Meyer, used for the first time a quicker and better equipped IBM 360 mainframe computer to do statistical analyses on survey data collected during the Detroit riots. Meyer was able to show that not only less educated Southerners were participating in the riots but also people who attended college. This story, published on the *Detroit Free Press*, won him a Pulitzer Prize together with other journalists and marked a paradigm shift in computer-assisted reporting. On the grounds of this success, Meyer not only supported the use of computers in journalistic practices but developed a whole new approach to investigative reporting by introducing and using social science research methods in journalism for data gathering, sampling, analysis, and presentation. In 1973 he published his thoughts in the seminal book entitled "Precision Journalism." The fact that computer-assisted reporting entered newsrooms especially in the USA was also

revealed through the increased use of computers in news organizations. In 1986, the *Time* magazine wrote that computers are revolutionizing investigative journalism. By trying to analyze larger databases, journalists were able to offer a broader perspective and much more information about the context of specific events.

The practice of computer-assisted reporting spread further until, at the beginning of the 1990s, it became a standard routine particularly in bigger newsrooms. The use of computers, together with the application of social science methods, has helped – according to Philip Meyer – to make journalism scientific. Besides, Meyer's approach tried also to tackle some of the common shortcomings of journalism like the increasing dependence on press releases, shrinking accuracy and trust, or the critique of political bias. An important factor of precision journalism was therefore the introduction and the use of statistical software. These programs enabled journalists for the first time to analyze bigger databases such as surveys or public records. This new approach might also be seen as a reaction to alternative journalistic trends that came up in the 1990s, for instance, the concept of new journalism. While precision journalism stood for scientific rigor in data analysis and reporting, new journalism used techniques from fiction to enhance reading experience.

There are some similarities between data journalism and computer-assisted reporting: both rely on specific software programs that enable journalists to transform raw data into news stories. However, there are also differences between computer-assisted reporting and data journalism, which are due to the context in which the two practices were developed. Computer-assisted reporting tried to introduce both informatics and scientific methods into journalism, given that at the time, data was scarce, and many journalists had to generate their own data. The rise of the Internet and new media contributed to the massive expansion of archives, databases, and to the creation of big data. There is no longer a poverty of information, data is now available in abundance. Therefore, data journalism is less about the creation of new databases, but more about data gathering, analysis, and

visualization, which means that journalists have to look for specific patterns within the data rather than merely seeking information – although recent discussions call for journalists to create their own databases due to an overreliance on public databases. Either way, the success of data journalism also led to new practices, routines, and mixed teams of journalists working together with programmers, developers, and designers within the same newsrooms, allowing them to tell stories in a different and visually engaging way.

Media Organizations and Big Data

Big data is not only a valuable resource for data journalism. Media organizations are data gatherers as well. Many media products, whether news or entertainment, are financed through advertising. In order to satisfy the advertisers' interests in the site's audience, penetration, and visits, media organizations track user behavior on their webpages. Very often, media organizations share this data with external research bodies, which then try to use the data on their behalf. Gathering information about their customers is therefore not only an issue when it comes to the use of social media. Traditional media organizations are also collecting data about their clients.

However, media organizations track the user behavior on news websites not only to provide data to their advertisers. Through user data, they also adapt the website's content to the audience's demand, with dysfunctional consequences for journalism and its democratic function within society. Due to web analytics and the generation of large-scale data collections, the audience exerts an increasing influence over the news selection process. This means that journalists – particularly in the online realm – are at the risk of increasingly adapting their news selections on the audience's feedback through data generated via web analytics. Due to the grim financial situation and their shrinking advertising revenue, some print media organizations especially in western societies try to apply strategies to compensate these deficits through a dominant market-driven discourse, manufacturing cheaper content that appeals to broader masses – publishing more soft news, sensationalism, and articles of human interest without

any connection to public policy issues. This is also due to the different competitive environment: while there are fewer competitors in traditional newspaper or broadcast markets, in the online world, the next competitor is just one click away.

Legacy media organizations, particularly newspapers and their online webpages, offer more soft news to increase traffic, to attract the attention of more readers, and thus to keep their advertisers at it. A growing body of literature about the consequences of this behavior shows that journalists, in general, are becoming much more aware of the audiences' preferences. At the same time, however, there is also a growing concern among journalists with regard to their professional ethics and the consequences for the function of journalism in society if they base their editorial decision-making processes on real-time data. The results of web analytics not only influence the placement of news on the websites; they also have an impact on the journalists' beliefs about what the audience wants. Particularly in online journalism, the news selection is carried out grounding the decisions on data generated by web analytics and no longer on intrinsic notions such as news values or personal beliefs. Consequently, online journalism becomes highly responsive to the audiences' preferences – serving less what would be in the public interest. As many news outlets are integrated organizations, which means that they apply a crossmedia strategy by joining previously separated newsrooms such as the online and the print staff, it might be possible that factors like data-based audience feedback will also affect print newsrooms. As Tandoc Jr. and Thomas state, if journalism continues to view itself as a sort of “conduit through which transient audience preferences are satisfied, then it is no journalism worth bearing the name” (Tandoc and Thomas 2015, p. 253).

While news organizations still struggle with self-gathered data due to the conflicts that can arise in journalism, media organizations active in the entertainment industry rely much more strongly on data about their audiences. Through large amounts of data, entertainment media can collect significant information about the audience's preferences for a TV series or a

movie – even before it is broadcast. Particularly for big production companies or film studios it is essential to observe structured data like ratings, market share, and box office stats. But also unstructured data like comments or videos in social media are equally important in order to understand consumer habits, given that they provide information about the potential success or failure of a (new) product.

An example of such use of big data is the launch of the TV show “House of Cards” by the Internet-based on demand streaming provider Netflix. Before launching the first original content with the political drama, Netflix was already collecting huge amounts of data about the streaming habits of their customers. Of more than 25 million users, they tracked around 30 million views a day (recording also when people are pausing, rewinding, or fast-forwarding the videos), about four million ratings, and three million searches (Carr 2013). On top of that, they also try to gather unstructured data from social media, and they look how customers are tagging the selected videos with metadata descriptors and whether they recommend the content. Based on these data, Netflix predicted possible preferences and decided to buy “House of Cards.” It was a major success for the online-based company.

There are also potential risks associated with the collection of such huge amounts of data: Netflix recommends specific movies or TV shows to their customers based on what they liked or what they have watched before. These recommendation algorithms might well guide the user toward more of their original content, without taking into account the consumers' actual preferences. In addition, consumers might not be able to discover new TV shows that transcend their usual taste. Given that services like Netflix know so much about their users' habits, another concern with regard to privacy arises.

Big Data Between Social Media, Ethics, and Surveillance

Social media are a main source for big data. Since the first major social media webpages have been launched in the 2000s, they began to collect and store massive amounts of data. These sites started

to gather information about the behavior, preferences, and interests of their users in order to know how their users would both think and act. In general, this process of datafication is used to target and tailor the services better to the users' interests. At the same time, social media use these large-scale data collections to help advertiser target the users. Big data in social media have therefore also a strong commercial connotation. Facebook's business model, for instance, is entirely based on hyper-targeted display ads. While display ads are a relatively old-fashioned way of addressing customers, Facebook can make it up with its incredible precision about the customers' interests and its ability to target advertising more effectively.

Big data are an integrative part of social media's business model: they possess far more information on their customers given that they have access not only to their surf behavior but above all to their tastes, interests, and networks. This might not only bear the potential to predict the users' behavior but also to influence it, particularly as social media such as Facebook and Twitter adapt also their noncommercial content to the individual users: the news streams we see on our personal pages are balanced by various variables (differing between social media) such as interactions, posting habits, popularity, the number of friends, user engagement, and others, being however constantly recombined. Through such opaque algorithms, social media might well use their own data to model voters: in 2010, for example, 61 million users in the USA were shown a banner message on their pages about how many of their friends already voted for the US Congressional Elections. The study showed that the banner convinced more than 340,000 additional people to cast their vote (Bond et al. 2012). The individually tailored and modeled messaging does not only bear the potential to harm the civic discourse; it also enhances the negative effects deriving from "asymmetry and secrecy built into this mode of computational politics" (Tufekci 2014).

The amount of data stored on social media will continue to rise, and already today, social media are among the largest data repositories in the world. Since the data collecting mania of social media will not decrease, which is also due to the

explorative focus of big data, it raises issues with regard to the specific purpose of the data collection. Particularly if the data usage, storage, and transfer remain opaque and are not made transparent, the data collection might be disproportionate. Yet, certain social media allow third parties to access their data, particularly as the trade of data increases because of its economic potential. This policy raises ethical issues with regard to transparency about data protection and privacy.

Particularly in the wake of the Snowden revelations, it has been shown that opaque algorithms and big data practices are increasingly important to surveillance: "[...] Big Data practices are skewing surveillance even more towards a reliance on technological "solutions," and that these both privileges organizations, large and small, whether public or private, reinforce the shift in emphasis toward control rather than discipline and rely increasingly on predictive analytics to anticipate and preempt" (Lyon 2014, p. 10). Overall, the Snowden disclosures have demonstrated that surveillance is no longer limited to traditional instruments in the Orwellian sense but have become ubiquitous and overly reliant on practices of big data – as governmental agencies such as the NSA and GCHQ are allowed to access not only the data of social media and search giants but also to track and monitor telecommunications of almost every individual in the world. However, the big issue even with the collect-all approach is that data is subject to limitations and bias, particularly if they rely on automated data analysis: "Without those biases and limitations being understood and outlined, misinterpretation is the result" (Boyd and Crawford 2012, p. 668). This might well lead to false accusation or failure of predictive surveillance as could be seen in the case of the Boston Marathon bombing case: first, a picture of the wrong suspect was massively shared on social media, and second, the predictive radar grounded on data gathering was ineffective.

In addition, the use of big data generated by social media entails also ethical issues in reference to scientific research. Normally, when human beings are involved in research, strict ethical rules, such as the informed consent of the people participating in the study, have to be observed. Moreover, in social media there are

“public” and “private,” which can be accessed. An example of such a controversial use of big data is a study carried out by Kramera et al. (2014). The authors deliberately changed the newsfeed of Facebook users: some got more happy news, others more sad ones, because the goal of the study was to investigate whether emotional shifts in those surrounding us – in this case virtually – can change our own moods as well. The issue with the study was that the users in the sample were not aware that their newsfeed was altered. This study shows that the use of big data generated in social media can entail ethical issues, not the least because the constructed reality within Facebook can be distorted. Ethical questions with regard to media and big data are thus highly relevant in our society, given that both the privacy of citizens and the protection of their data are at stake.

Conclusion

Big data plays a crucial role in the context of the media. The instruments of computer-assisted reporting and data journalism allow news organizations to engage in new forms of investigations and storytelling. Big data also allow media organizations to better adapt their services to the preferences of their users. While in the news business this may lead to an increase of soft news, the entertainment industry benefits from such data in order to predict the audience’s taste with regard to potential TV shows or movies. One of the biggest issues with regard to media and big data are its ethical implications, particularly with regard to data collection, storage, transfer, and surveillance. As long as the urge to collect large amounts of data and the use of opaque algorithms continue to prevail in many already powerful (social) media organizations, the risks of data manipulation and modeling will increase, particularly as big data are becoming even more important in many different aspects of our lives. Furthermore, as the Snowden revelations showed, collect-it-all surveillance already relies heavily on big data practices. It is therefore necessary to increase both the research into and the awareness about the ethical implications of big data in the media context. Only thanks

to a critical discourse about the use of big data in our society, we will be able to determine “our agency with respect to big data that is generated *by* us and *about* us, but is increasingly being used *at* us” (Tufekci 2014). Being more transparent, accountable, and less opaque about the use and, in particular, the purpose of data collection might be a good starting point.

Cross-References

- Crowdsourcing
- Digital Storytelling, Big Data Storytelling
- Online Advertising
- Transparency

References

- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489, 295–298.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662–679.
- Carr, D. (2013, February 24). Giving readers what they want. *New York Times*. <http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html>. Accessed 11 July 2016.
- Kramera, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24), 8788–8790.
- Lyon, D. (2014, July–December). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society*, 1–13.
- Tandoc Jr., E. C., & Thomas, R. J. (2015). The ethics of web analytics. Implications of using audience metrics in news construction. *Digital Journalism*, 3(2), 243–258.
- Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*, 19(7). <http://journals.uic.edu/ojs/index.php/fm/article/view/4901/4097>. Accessed 12 July 2016.

Media Ethics

- Media

Medicaid

Kim Lorber¹ and Adele Weiner²

¹Social Work Convening Group, Ramapo College of New Jersey, Mahwah, NJ, USA

²Audrey Cohen School For Human Services and Education, Metropolitan College of New York, New York, NY, USA

Introduction

Medicaid provides medical care to low-income individuals and is the US federal government's most costly welfare program. Funds are provided to states that wish to participate and programs are managed differently within each; all states have participated since 1982. As of January 2017, 32 states participate in the Medicaid expansion under the Affordable Care Act (ACA). In 2015, according to the Kaiser Family Foundation (KFF), 20% of Americans with medical insurance, or over 62 million people, were covered by Medicaid. By 2017, with the ACA expansion, Medicaid has become the nation's largest insurer, covering over 74.5 million people or 1 in 5 individuals and has financed over 16% of all personal health care. A poll in 2017 found that 50% of respondents said Medicaid is important to their family (Kaiser Family Foundation 2017). It is estimated that the proposed alternatives to the ACA will throw 24 million people off Medicaid in the next 10 years.

Medicaid data is big by nature of the number of individuals served and information collected. The potential for understanding the population served regarding medical, hospital, acute and long-term care is seemingly limitless. And, with such potential comes huge responsibility to be comprehensive and accurate.

Privacy and Health Care Coordination

The Health Information and Portability Act of 1966 (HIPAA) provides safeguards to ensure privacy of medical records, including any

information generated for those covered by Medicaid. It provides national standards for processing electronic healthcare transactions, including secure electronic access to health data. Electronic health networks or health information exchanges (HIE) facilitate the availability of medical information electronically across organizations within states, regions, communities, or hospital systems. Such systems include medical records for all patients, including their health insurance information.

Data for Medicaid enrollment, service utilization and expenditures, is collected by the Medicaid and Statistical Information System (MSIS) used by the Centers for Medicare and Medicaid Services (CMS). Ongoing efforts continue to ensure the availability of necessary medical data for providers, reduce duplication of services, and to insure timely payment. In 2014, CMS implemented the Medicaid Innovation Accelerator Program (IAP) to improve health care for Medicaid beneficiaries by supporting individual states' efforts in reducing costs while improving payment and delivery systems. In 2016, CMS created an interoperability initiative to connect a wider variety of Medicaid providers and improve health information exchanges.

Ultimately, it is possible to track Medicaid beneficiaries in real-time as they utilize the health care system in any environment ensuring appropriate treatment in the most cost-effective venues. These real-time details allow Medicaid providers throughout the country to access detailed reports about patient treatment as well as program spending to managed care plans which do not necessarily use a fee-for-service system but, instead, make the provider fully responsible for potentially high quality and costly treatments.

Reducing Medicaid Costs

A variety of methods are being utilized throughout the country in order to reduce Medicaid costs, increase efficiency, demonstrate billing accountability, and find cases of Medicaid fraud. In a volatile and politically charged healthcare environment, Medicaid services and eligibility are

constantly changing. Sorting through double billing, patients' repeat visits, and the absence of required follow-ups have demonstrated the benefits in cost-saving measures such as preventing potential hospitalizations while ensuring proper patient treatment and follow up. Individual states are responsible for developing electronic systems for managing their Medicaid costs.

Washington State, facing a fiscal Medicaid crisis, implemented a statewide database of ER visits, available across hospitals, to document and reduce the use of hospitals for non-emergency care. While hospitals cannot turn away a patient, reimbursement for treatment may not be provided due to limited Medicaid funds. The state had tried different approaches including limiting the number of ER visits per year to identifying 500 ailments that would no longer be reimbursable as emergency care. The solution to these protested and rejected "solutions" was to create a database that, within minutes of arrival, allows an attending doctor to see a patient's complete medical history, potentially reducing duplication of diagnostic tests. Referrals to alternate more appropriate and less expensive treatment resources resulted in a Medicaid cost reduction in the course of 1 year of \$33.7 million in part attributable to the database. Oregon will follow and inquiries from other states to reduce ER costs have been received as Washington has shown how impactful data can be.

IBM developed text analysis software, which was successful in reducing Medicaid re-admissions in North Carolina. Other uses have resulted in systems which provide alerts to case managers and others to remind patients to follow up with specialists or to complete necessary medical tests in order to complete the treatment begun in the hospital. Such efforts allow patients to truly address their malady and reduce hospital readmissions while resulting in reduced Medicaid costs.

The Medicaid Management Information System (MMIS) developed by Xerox and, approved by the Centers for Medicare and Medicaid Services (CMS), is used in at least 31 states. The sophisticated algorithms review drug prescriptions, pharmacy and doctors' offices' repeated

rule violations, and find duplicate billing and reused prescriptions resulting in multiple and fraudulent payments. Advances are allowing predictive detection of fraud to avoid only finding it using the rules-based system and after payments have been made. This helps defeat criminal activities as these are enhanced in synch with regularly improved preventive methods.

Limitations of Medicaid Big Data

With Big Data comes big responsibility. All data collection networks must be HIPAA compliant and protect patient medical information and yet must be accessible to service providers. Biola et al. (2014) analyzed Medicaid information from North Carolina. The study was of non-cancer adults on Medicaid who had received at least 10 computed tomography (CT) scans, to inform them of their radiation exposure. Most interesting as relevant to this chapter is that scan information was *only* available about Medicaid patients and even that was not comprehensive as some patients with high exposure were excluded unintentionally because care providers' file claims differed by setting. Thus Medicaid information can be incomplete suggesting the need for future alignment of billing and claims systems.

Conclusion

Big data, as related to Medicaid, can significantly improve patient safety and care while providing cost-saving measures. As political challenges are mounted to the Affordable Care Act, Medicaid data may help to inform the national discussion about health and insurance. It clearly demonstrates how access to health care can reduce more costly emergency room visits. Demographic information about those enrolled in Medicaid and their advocates can present a sizeable voting block in the political process to protect Medicaid funding levels and eligibility for enrollment by highlighting efficiencies in Medicaid payment and services.

Further Reading

- Biola, H., Best, R. M., Lahlou, R. M., Burke, L. M., Dward, C., Jackson, C. T., Broder, J., Grey, L., Semelka, R. C., & Dobson, A. (2014). With “big data” comes big responsibility: Outreach to North Carolina Medicaid patients with 10 or more computed typography scans in 12 months. *North Carolina Medical Journal*, 75(2), 102–109.
- Kaiser Family Foundation. (2017). Medicaid. Retrieved on May 13, 2017 from <http://kff.org/medicaid/>.
- Medicaid.gov: Keeping American Healthy. (2014). Retrieved on September 20, 2014 from <http://www.medicaid.gov/>.
- Weise, K. (2014, March 25). How big data helped cut emergency room visits by 10 percent. Retrieved on September 10, 2014 from <http://www.businessweek.com/articles/2014-03-25/how-big-data-helped-cut-emergency-room-visits-by-10-percent>.

Metadata

Xiaogang Ma

Department of Computer Science, University of Idaho, Moscow, ID, USA

Metadata are data about data, or in a more general sense, they are data about resources. They provide a snapshot about a resource, such as information about the creator, date, subject, location, time and methods used, etc. There are high-level metadata standards that can provide a general description of a resource. In recent years, community efforts have been taken to develop domain-specific metadata schemas and encode the schemas with machine readable formats for the World Wide Web. Those schemas can be reused and extended to fit requirements of specific applications. Comparing with the long-term archive of data and metadata in traditional data management and analysis, the velocity of Big Data leads to short-term and quick applications addressing scientific and business issues. Accordingly, there is a metadata data life cycle in Big Data applications. Community metadata standards and machine readable formats will be a big advantage to facilitate the metadata data life cycle on the Web.

Know Before Use

Few people are able to use a piece of data before knowing its subject, origin, structure, and meaning. A primary functionality of metadata is to help people to obtain an overview of some data, and this functionality can be understood through a few real-world examples. If data are comparable with goods in a grocery, then metadata are like the information on the package of an item. A consumer may care more about the ingredients due to allergies to some substances, the nutrition facts due to dietary needs, and/or the manufacturer and date of expiration due to personal preferences. Most people want to know the information about a grocery item before purchasing and consuming it. The information on the package provides a concise and essential introduction about the item inside. Such nutrition and ingredient information of grocery items is mandatory for manufacturers in many countries. Similarly, an ideal situation for data users is that they can receive clear metadata from data providers. However, compared to the food industry, the rules and guidelines for metadata are still less developed.

Another comparable subject is the 5W1H method for storytelling or context description, especially in journalism. The 5W1H represents the question words who, what, when, where, why, and how, which can be used to organize a number of questions about a certain object or event, such as: Who is responsible for a research project? What are the planned output data? Where will the data be archived? When will the data be open access? Why a specific instrument is needed for data collection? How will the data be maintained and updated? In journalism, the 5W1H is often used to evaluate whether the information covered in a news article is complete or not. Normally, the first paragraph of a news article gives a brief overview of the article and provides concise information to answer the 5W1H questions. By reading the first paragraph, a reader can grasp the key information of an article even before reading through the full text. Metadata is data about data; such functionality is similar to what the first paragraph works for a news article, and

metadata items used for describing a dataset are equal to the 5W1H question words.

Metadata Hierarchy

Metadata are used for describing resources. The description can be general or detailed according to the actual needs. Accordingly, there is a hierarchy of metadata items corresponding to the actual needs of describing an object. For instance, the abovementioned 5W1H question words can be regarded as a list of general metadata items, and they can also be used to describe datasets. However, the six question words only offer a start point, and there may be various derived metadata items in actual works. In early days there was such a heterogeneous situation among the metadata provided by different stakeholders. To promote standardization of metadata items, a number of international standards have been developed. The most well-known standard is the Dublin Core Metadata Element Set (DCMI Usage Board 2012). The name “Dublin” originates from a 1995 workshop at Dublin, OH, USA. The word “Core” means that the elements are generic and broad. The 15 core elements are contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type. Those elements are more specific than the 5W1H question words and can be used for describing a wide range of resources, including datasets. The Dublin Core Metadata Element Set was published as a standard by the International Organization for Standardization (ISO) in 2003 and later revised in 2009. It has also been endorsed by a number of other national or international organizations such as the American National Standards Institute and the Internet Engineering Task Force.

The 15 core elements are part of an enriched specification of metadata terms maintained by the Dublin Core Metadata Initiative (DCMI). The specification includes properties in the core elements, properties in an enriched list of terms, vocabulary encoding schemes, syntax encoding schemes, and classes (including the DCMI Type Vocabulary). The enriched terms include all the

15 core elements and cover a number of more specific properties, such as abstract, access rights, has part, has version, medium, modified, spatial, temporal, valid, etc. In practice, the metadata terms in the DCMI specification can be further extended by combining with other compatible vocabularies to support various application profiles. With the 15 core elements, one is able to provide rich metadata for a certain resource, and by using the enriched DCMI metadata terms and external vocabularies, one can create an even more specific metadata description for the same object. This can be done in a few ways. For example, one way is to use terms that are not included in the core elements, such as spatial and temporal. Another possible way is to use a refined metadata term that is more appropriate for describing an object. For instance, the term “description” in the core elements is with broad meaning, and it may include an abstract, a table of contents, a graphical representation, or a free-text account of a resource. In the enriched DCMI terms, there is a more specific term “abstract,” which means a summary of a resource. Compared to “description,” the term “abstract” is more specific and appropriate if one wants to collect a literal summary of an academic article.

Domain-Specific Metadata Schemas

High-level metadata terms such as those in the Dublin Core Metadata Element Set have broad meaning and are applicable to various resources. However, those metadata elements are too general in meaning and sometimes are implicit. If one wants a more specific and detailed description of the resources, a domain-specific metadata schema is needed. Such a metadata schema is a list of organized metadata items for describing a certain type of resource. For example, there could be a metadata schema for each type defined in the DCMI Type Vocabulary, such as dataset, event, image, physical object, service, etc. There have been various national and international community efforts for building domain-specific metadata schemas. Especially, many schemas developed in recent years face the data management and

exchange on the Web. A few recent works are introduced below.

The data catalog vocabulary (DCAT) (Erickson and Maali 2014) was approved as a World Wide Web Consortium (W3C) recommendation in January 2014. It was designed to facilitate interoperability among data catalogs published on the Web. DCAT defines a metadata schema and provides a number of examples on how to use it. DCAT reuses a number of DCMI metadata terms in combination with terms from other schemas such as the W3C Simple Knowledge Organization System (SKOS). It also defines a few new terms to make the resulted schema more appropriate for describing datasets in data catalogs.

The Darwin Core is a group of standards for biodiversity applications. By extending the Dublin Core metadata elements, the Darwin Core establishes a vocabulary of terms to facilitate the description and exchange of data about the geographic occurrence of organisms and the physical existence of biotic specimens. The Darwin Core itself is also extensible, which provides a mechanism for describing and sharing additional information.

The ecological metadata language (EML) is a metadata standard developed for the none-geospatial datasets in the field of ecology. It is a set of schemas encoded in the format of extensible markup language (XML) and thus allows structured expression of metadata. EML can be used to describe digital resources and also nondigital resources such as paper maps.

The international geo sample number (IGSN), initiated in 2004, is a sample identification code for the geoscience community. Each registered IGSN identifier is accompanied with a group of metadata providing detailed background information about that sample. Top concepts in the current IGSN metadata schema are sample number, registrant, related resource identifiers, and log. A top concept may include a few child concepts. For example, there are two child concepts for “registrant”: registrant name and name identifier.

The ISO 19115 and ISO 19115-2 geographic information metadata are regarded as a best practice of metadata schemas for geospatial data.

Geospatial data are about objects with some position on the surface of the Earth. The ISO 19115 standards provide guidelines on how to describe geographical information and services. Detailed metadata items cover topics about contents, spatiotemporal extents, data quality, channels for access and rights to use, etc. Another standard, ISO 19139, provides an XML schema implementation for the ISO 19115. The catalog service for the Web (CSW) is an open geospatial consortium (OGC) standard for describing online geospatial data and services. It adopts ISO 19139, the Dublin Core elements and items from other metadata efforts. Core elements in CSW include title, format, type, bounding box, coordinate reference system, and association.

Annotating a Web of Data

Recent efforts on metadata standards and schemas, such as the abovementioned Dublin Core, DCAT, Darwin Core, EML, IGSN metadata, ISO 19139, and CSW, show a trend of publishing metadata on the Web. More importantly, by using standard encoding formats, such as the XML and W3C resource description framework (RDF), they are making metadata machine discoverable and readable. This mechanism moves the burden of searching, evaluating, and integrating massive datasets from humans to computers, and for computers such burden is not real burden because they can find ways to access various data sources through standardized metadata on the Web. For example, the project OneGeology aims to enable online access to geological maps across the world. By the end of 2014, the OneGeology has 119 participating nations, and most of them share national or regional geological maps through OGC geospatial data service standards. Those map services are maintained by their corresponding organizations, and they also enable standardized metadata services, such as CSW. On the one hand, OneGeology provides technical supports to organizations who want to set up geologic map services using common standards. On the other hand, it also provides a central data portal for end users to access various distributed

metadata and data services. The OneGeology project presents a successful example on how to rescue the legacy data, update them with well-organized metadata, and make them discoverable, accessible, and usable on the Web.

Comparing with domain-specific structured datasets, such as those in OneGeology, many other datasets in the Big Data are not structured, such as webpages and data stream on social media. In 2011, the search engines Bing, Google, Yahoo!, and Yandex launched an initiative called schema.org, which aims at creating and supporting a common set of schemas for structured data markup on web pages. The schemas are presented as lists of tags in hypertext markup language (HTML). Webmasters can use those tags to mark up their web pages, and search engine spiders and other parsers can recognize those tags and record what a web page is about. This makes it easier for search engine users to find the right web pages. Schema.org adopts a hierarchy to organize the schemas and vocabularies of terms. The concept on the top is thing, which is very generic and is divided into schemas of a number of child concepts, including creative work, event, intangible, medical entity, organization, person, place, product, and review. These schemas are further divided into smaller schemas with specific properties. A child concept inherits characteristics from a parent concept. For example, book is a child concept of creative work. The hierarchy of concepts and properties does not intend to be a comprehensive model that covers everything in the world. The current version of schema.org only represents those entities that the search engines can handle in a short term. Schema.org provides a mechanism for extending the scope of concepts, properties, and schemas. Webmasters and developers can define their own specific concepts, properties, and schemas. Once those extensions are commonly used on the Web, they can also be included as a part of the schema.org schemas.

Linking for Tracking

If the recognition of domain-specific topics is a work to identify resource types, then the definition

of metadata items is a work of annotating those types. The work in schema.org is an excellent reflection of those two works. Various structured and unstructured resources can be categorized and annotated by using metadata and are ready to be discovered and accessed. In a scientific or business procedure, various resources are retrieved and used, and outputs are generated and archived and perhaps be reused elsewhere. In recent years, people take a further step to make links among those resources, their types, and properties, as well as the people and activities involved in the generation of those outputs. The work of categorization, annotation, and linking as a whole can be used to describe the origin of a resource, which is called provenance. There have been community efforts developing specifications of commonly usable provenance models.

The Open Provenance Model was initiated in 2006. It includes three top classes: artifact, process, and agent and their subclasses, as well as a group of properties, such as was generated by, was controlled by, was derived from, and used, for describing the classes and the interrelationships among them. Another earlier effort is the proof markup language, which was used to represent knowledge about how information on the Web was asserted or inferred from other information sources by intelligent agents. Information, inference step/inference rule, and inference engine are the three key building blocks in the proof markup language.

Works on the Open Provenance Model and the proof markup language have set up the basis for community actions. Most recently, the W3C approved the PROV Data Model as a recommendation in 2013. The PROV Data Model is a generic model for provenance, which allows specific representations of provenance in research domains or applications to be translated into the model and be interchangeable among systems (Moreau and Missier 2013). There are intelligent knowledge systems that can import the provenance information from multiple sources, process it, and reason over it to generate clues for potential new findings. The PROV Data Model includes three core classes, entity, activity, and agent, which are comparable to the Open Provenance

Model and the proof markup language. W3C also approved the PROV Ontology as a recommendation for the expression of the PROV Data Model with semantic Web languages. It can be used to represent machine readable provenance information and can also be specialized to create new classes and properties to represent provenance information of specific applications and domain. The extension and specification here are similar to the idea of a metadata hierarchy. A typical application of the PROV Ontology is the Global Change Information System for the US Global Change Research Program (Ma et al. 2014), which captures and presents provenance of global change research, and links to the publications, datasets, instruments, models, algorithms, and workflows that support key research findings. The provenance information in the system increases understanding, credibility, and trust in the works of the US Global Change Research Program and aids in fostering reproducibility of results and conclusions.

A Metadata Life Cycle

Velocity is a unique feature that differentiates Big Data from traditional data. Traditional data can also be big, but they have a relatively longer life cycle compared to social media data stream in Big Data. Big Data life cycles are featured by short-term and quick deployments to solve specific scientific or business issues. In traditional data management, especially for a single data center or data repository, the metadata life cycle is less addressed. Now, facing the short-lived and quick Big Data life cycles, attention should also be paid to the metadata life cycle.

In general, a data life cycle covers steps of context recognition, data discovery, data access, data management, data archive, and data distribution. Correspondingly, a metadata life cycle covers similar steps but they focus on the description of data rather than the data themselves. The context recognition allows people to study a specific domain or application and reuse any existing metadata standards and schemas. Then in the metadata discovery step, it is possible to develop

applications to automatically harvest machine readable metadata from multiple sources and harmonize them. Commonly used domain-specific metadata standards and machine readable formats will significantly facilitate the metadata life cycle in applications using Big Data, because most of such applications will be on the Web and interchangeable schemas and formats will be an advantage.

Cross-References

- [Data Brokers and Data Services](#)
- [Data Profiling](#)
- [Data Provenance](#)
- [Data Sharing](#)
- [Open Data](#)

Further Reading

- DCMI Usage Board. (2012). *DCMI metadata terms*. <http://dublincore.org/documents/dcmi-terms>.
- Erickson, J., Maali, F. (2014). *Data catalog vocabulary (DCAT)*. <http://www.w3.org/TR/vocab-dcat>.
- Ma, X., Fox, P., Tilmes, C., Jacobs, K., & Waple, A. (2014). Capturing provenance of global change information. *Nature Climate Change*, 4/6, 409–413.
- Moreau, L., Missier, P.. (2013). *PROV-DM: The PROV data model*. <http://www.w3.org/TR/prov-dm>.

Middle East

Feras A. Batarseh
College of Science, George Mason University,
Fairfax, VA, USA

Synonyms

[Mid-East](#); [Middle East and North Africa \(MENA\)](#)

Definition

The Middle East is a transcontinental region in Western Asia and North Africa. Countries of the

Middle East are ones extending from the shores of the Mediterranean Sea, south towards Africa, and east towards Asia, and sometimes beyond depending on the context (political, geographical, etc.). The majority of the countries of the region speak Arabic.

Introduction

The term “Middle East” evolved with time. It was originally referred to as the countries of the Ottoman empire, but by the mid-twentieth century, a more common definition of the Middle East included the following states (countries): Turkey, Jordan, Cyprus, Lebanon, Iraq, Syria, Israel, Iran, the West Bank and the Gaza Strip (Palestine), Egypt, Sudan, Libya, Saudi Arabia, Kuwait, Yemen, Oman, Bahrain, Qatar, and United Arab Emirates (UAE). Subsequent political and historical events have tended to include more countries into the mix (such as: Tunisia, Algeria, Morocco, Afghanistan, and Pakistan).

The Middle East is often referred to as the cradle of civilization. By studying the history of the region, it is clear why the first human civilizations were established in this part of the world (particularly the Mesopotamia region around the Tigris and Euphrates rivers). The Middle East is where humans made their first transitions from nomadic to agriculture, invented the wheel, created basic agriculture, and where the beginnings of the written-word first existed. It is well known that this region is an active political, economic, historic, and religious part of the world (Encyclopedia Britannica 2017). For the purposes of this encyclopedia, the focus of this entry is on technology, data, and software of the Middle East.

The Digital Age in the Middle East

Since the beginning of the 2000s, the Middle East was one of the highest regions in the world in terms of adoption of social media; certain countries (such as the United Arab Emirates, Qatar, and

Bahrain) have adopted social technologies by 70% of its population (which is a higher percentage than the United States). While citizens are jumping on the wagon of social media, governments still struggle to manage, define, or guide the usage of such technologies.

The *McKinsey Middle East Digitization Index* is the one of the main metrics to assess the level and impact of digitization across the Middle East. Only 6% of Middle Eastern public lives under a digitized smart or electronic government (The UAE, Jordan, Israel, and Saudi Arabia are among the few countries that have some form of e-government) (Elmasri et al. 2016). However, many new technology startups are coming from the Middle East with great success. The most famous technology startup companies coming out of the Middle East include: (1) Maktoob (from Jordan): is one that stands out. The company represents a major trophy on the list of Middle Eastern tech achievements. It made global headlines when it was bought by Yahoo, Inc. for \$80 million in 2009, symbolizing a worldwide important step by a purely Middle Eastern company. (2) Yamli (from Lebanon): One of the most popular web apps for Arabic speakers today. (3) GetYou (from Israel): A famous social media application. (4) Digikala (from Iran): An online retailer application. (5) ElWafeyat (from Egypt): An Arabic language social media site for honoring deceased friends and family. (6) Project X (from Jordan): A mobile application that allows for 3D printing of prosthetics, inspired by wars in the region. These examples are assembled from multiple sources; many other exciting projects exist as well (such as Souq which was acquired by Amazon in 2017, Masdar, Namshi, Sukar, and many others).

Software Arabization: The Next Frontier

The first step towards invoking more technology in a region is to localize the software, content, and its data. Localizing a software system is accomplished by supporting a new spoken language (Arabic Language in this context, hence the name, Arabization). A new term is presented in

this entry of the Encyclopedia, **Arabization**: it is the overall concept that includes the process of making the software available and reliable across the geographical borders of the Arab states. Different spoken languages have different orientations and fall into different groups. Dealing with these groups is accomplished by using different *code pages* and *Unicode fonts*. Languages fall into two main families, single-byte (such as: French, German, and Polish) and double-byte (such as: Japanese, Chinese, and Korean). Another categorization that is more relevant to Middle Eastern Languages is based on their orientation. Most Middle Eastern languages are right-to-left (RTL) (such as: Arabic and Hebrew), while other world languages are left-to-right (LTR) (such as: English and Spanish). For all languages, however, a set of translated strings should be saved in a *bundle file* that indexes all the strings, assign them IDs so the software program can locate them and display the right string in the language of the user. Furthermore, to accomplish software Arabization, characters encoding should be enabled. The default encoding for a given system is determined by the runtime locale set on the machine's operating system. The most commonplace character encoding format is UTF (USC transformation format) USC is the universal character set. UTF is designed to be compatible with ASCII. UTF has three types: UTF-8, UTF-16, and UTF-32. UTF is the international standard for ISO/IEC 10646. It is important to note that the process of Arabization is not a trivial process; engineers cannot merely inject translated language strings into the system, or hardcode cultural, date, or numerical settings into the software, rather, the process is done by obtaining different files based on the settings of the machine, the desires of the user, and applying the right locales. An Arabization package needs to be developed to further develop the digital, software, and technological evolution in the Middle East.

Bridging the Digital Divide

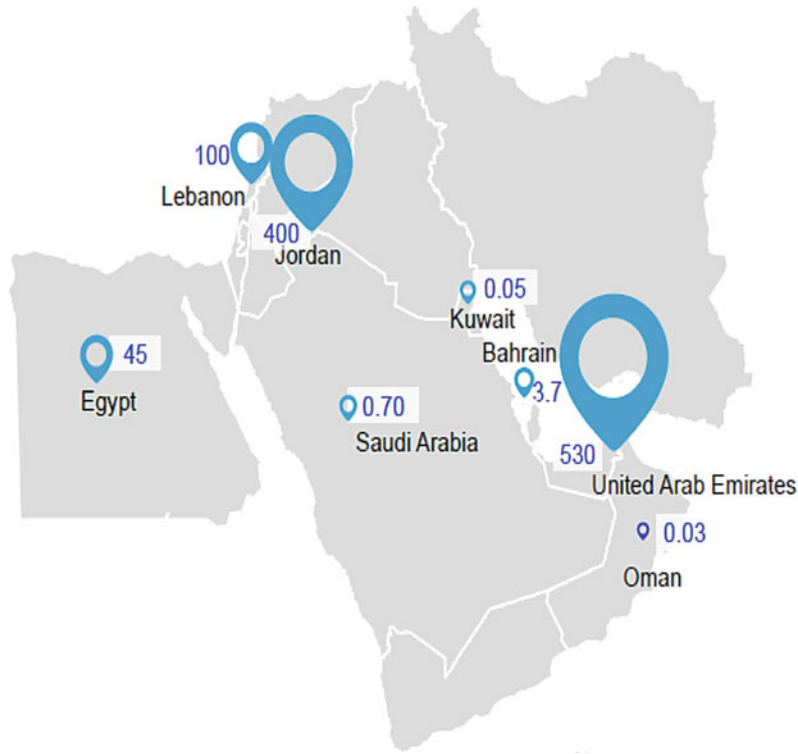
Information presented in this entry showed how the Middle East is speeding towards catching-

up with industrialized nations in terms of software technology adoption and utilizations (i.e., bridge the digital divide between third world and first world countries). Figure 1 below shows which countries are investing towards leading that transformation; numbers in the figure illustrate venture capital funding as share of GDP (Elmasri et al. 2016). However, According to Cisco's 2015 visual networking index (VNI), the world is looking towards a new digital divide, beyond software and mobile apps. By 2019, the number of people connecting to Internet is going to rise to 3.9 billion users, reaching over 50% of the global population. That will accelerate the new wave of big data, machine learning, and the Internet of Things (IoT). That will be the main new challenge for technology innovators in the Middle East. Middle Eastern countries need to first lay the "data" infrastructure (such as the principle of software Arabization presented above) that would enable the peoples of the Middle East towards higher adoption rates of future trends (big data and IoT). Such a shift would greatly influence economic growth at countries all across the region; however, the impacts of technology require minimum adoption thresholds before those impacts begin to materialize; the wider the intensity and use of big data, Internet of things (IoT), and machine learning, the greater the impacts.

Conclusion

The Middle East is known for many historical and political events, conflicts, and controversies; however, it is not often referred to as a technological and software-startup hub. This entry of the Encyclopedia presents a brief introduction to the Middle East and draws a simple picture about its digitization, and claims that Arabization of software could lead to many advancements across the region and eventually the world – for startups and creativity, the Middle East is an area worth watching (Forbes 2017).

Middle East,
Fig. 1 Middle Eastern
Investments in Technology
(Elmasri et al. 2016)



References

Elmasri, T., Benni, E., Patel, J., & Moore, J. (2016). *Digital Middle East: Transforming the region into a leading digital economy*. McKinsey and Company. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0ahUKEwiG2J2e55LTAhXoiVQKHfD8CxAQFggfMAE&url=http%3A%2F%2Fwww.mckinsey.com%2F~%2Fmedia%2Fmckinsey%2Fglobal%2520themes%2Fmiddle%2520east%2520and%2520africa%2Fdigital%2520middle%2520east%2520transforming%2520the%2520region%2520into%2520a%2520leading%2520digital%2520economy%2Fdigital-middle-east-finalupdated.ashx&usg=AFQjCNHioXhFY692mS_Qwa6hkBT6UiXYVg&sig2=6udbc7EP-bPs-ygQ18KSLA&cad=rja.

Encyclopedia Britannica. (2017). Available at <https://www.britannica.com/place/Middle-East>.

Forbes reports on the Middle East. (2017). Available at <http://www.forbes.com/sites/natalierobehmed/2013/08/22/forget-oil-tech-could-be-the-next-middle-east-goldmine/>.

Mid-East

- Middle East

Middle East and North Africa (MENA)

- Middle East

Mixture-of-Experts

- Ensemble Methods

Mobile Analytics

Ryan S. Eanes
Department of Business Management,
Washington College, Chestertown, MD, USA

Analytics, broadly defined, refers to a series of quantitative measures that allow marketers, vendors, business owners, advertisers, and interested

parties the ability to gauge consumer engagement and interaction with a property. When properly deployed and astutely analyzed, analytics can help to inform a range of business decisions related to user experience, advertising, budgets, marketing, product development, and more. Mobile analytics, then, refers to the measurement of consumer engagement with a brand, property, or product via a mobile platform, such as a smartphone or tablet computer.

Despite the fact that the mobile Internet and app markets have exploded in growth over the past decade, and despite the fact that more than half of all American adults now own at least one smartphone, according to the Pew Research Center, marketers have been relatively slow to jump into mobile marketing. In fact, American adults spend at least 20% of their time online via mobile devices; the advertising industry has been playing “catch-up” over the past few years in an attempt to chase this market. Even so, analyst Mary Meeker notes that advertising budgets still devote only about a tenth of their expenditures to mobile – though this is a fourfold increase from just a few years ago.

Any entity that is considering the deployment of a mobile strategy must understand consumer behavior as it occurs via mobile devices. Web usability experts have known for years that online browsing behavior can be casual, with people quickly clicking from one site to another and making judgments about content encountered in mere seconds. Mobile users, on the other hand, are far more deliberate in their efforts – generally speaking, a mobile user has a specific task in mind when he or she pulls out his phone. Browsing is far less likely to occur in a mobile context. This is due to a number of factors, including screen size, connection speed, and the environmental context in which mobile activity takes place – the middle of the grocery store dairy case, for example, is not the ideal place for one to contemplate the purchase of an eight-person spa for the backyard.

The appropriate route to the consumer must be considered, as well. This can be a daunting prospect, particularly for small businesses, businesses with limited IT resources, or businesses with little

previous web or tech experience. If a complete end-user experience is desired, there are two primary strategies that a company can employ: an all-in-one web-based solution, or a stand-alone app.

All-in-one web-based solutions allow the same HTML5/CSS3-based site to appear elegant and functional in a full-fledged computer-based browser while simultaneously “degrading” on a mobile device in such a way that no functionality is lost. In other words, the same underlying code provides the user experience regardless of what technological platform one uses to visit a site. There are several advantages to this approach, including singularity of platform (that is, no need to duplicate properties, logos, databases, etc.), ease of update, unified user experience, and relative ease of deployment. However, there are downsides: full implementation of HTML5 and CSS3 are relatively new. As a result, it can be costly to find a developer who is sufficiently knowledgeable to make the solution as seamless as desired, and who can articulate the solution in such a way that non-developers will understand the full vision of the end product. Furthermore, development of a polished finished product can be time-consuming and will likely involve a great deal of compromise from a design perspective.

Mobile analytics tools are relatively easy to deploy when a marketer chooses to take this route, as most modern smartphone web browsers are built on the same technologies that drive computer-based web browsers – in other words, most mobile browsers support both JavaScript and web “cookies,” both of which are typically requisites for analytics tools. Web pages can be “tagged” in such a way that mobile analytics can be measured, which will allow for the collection of a variety of information on visitors. This might include device type, browser identification, operating system, GPS location, screen resolution/size, and screen orientation, all of which can provide clues as to the contexts in which users are visiting the website on a mobile device. Some mainstream web analytics tools, such as Google Analytics, already include a certain degree of information pertaining to mobile users (i.e., it is possible to drill down into reports and determine

how many mobile users have visited and what types of devices they were using); however, marketing entities that want a greater degree of insight into the success of their mobile sites will likely need to seek out a third-party solution to monitor performance.

There are a number of providers of web-based analytics solutions that cover mobile web use. These include, but are not limited to, ClickTale, which offers mobile website optimization tools; comScore, which is known for its audience measurement metrics; Flurry, which focuses on use and engagement metrics; Google, which offers both free and enterprise-level services; IBM, which offers the ability to record user sessions and perform deep analysis on customer actions; Localytics, which offers real-time user tracking and messaging options; Medio, which touts “predictive” solutions that allow for custom content creation; and Webtrends, which incorporates other third-party (e.g., social media) data.

The other primary mobile option: development of a stand-alone smartphone or tablet app. Stand-alone apps are undeniably popular, given that 50 billion apps were downloaded from the Apple App Store between July 2008 and June 2014. A number of retailers have had great success with their apps, including Amazon, Target, Zappos, Groupon, and Walgreens, which speaks to the potential power of the app as a marketing tool. However, consider that there are more than one million apps in the Apple App Store alone, as of this writing – those odds greatly reduce the chances that an individual will simply “stumble across” a company’s app in the absence of some sort of viral advertising, breakout product, or buzzworthy word-of-mouth. Furthermore, developing a successful and enduring app can be quite expensive, particularly considering that a marketer will likely want to make versions of the app available for both Apple iOS and Google Android (the two platforms are incompatible with each other). Estimates for app development vary widely, from a few thousand dollars at the low end all the way up to six figures for a complex app, according to Mark Stetler of AppMuse – and these figures do not include ongoing updates, bug fixes, or recurring content updates, all of which

require staff with specialized training and know-how.

If a full-fledged app or redesigned website proves too daunting or beyond the scope of what a marketer needs or desires, there are a number of other techniques that can be used to reach consumers, including text and multimedia messaging, email messaging, mobile advertising, and so forth. Each of these techniques can reveal a wealth of data about consumers, so long as the appropriate analytic tools are deployed in advance of the launch of any particular campaign.

Mobile app analytics are quite different from web analytics in a number of ways, including the vocabulary. For example, there are no page views in the world of app analytics – instead, “screen views” are referenced. Likewise, an app “session” is analogous to a web “visit.” App analytics often have the ability to access and gauge the use of various features built into a phone or tablet, including the accelerometer, GPS, and gyroscope, which can provide interesting kinesthetic aspects to user experience considerations. App analytics tools are also typically able to record and retain data related to offline usage for transmission when a device has reconnected to the network, which can provide a breadth of environmentally contextual information to developers and marketers alike. Finally, multiple versions of a mobile app can exist “in the wild” simultaneously because users’ proclivities differ when it comes to updating apps. Most app analytic packages have the ability to determine which version of an app is in use so that a development team can track interactional differences between versions and confirm that bugs have been “squashed.”

As mentioned previously, marketers who choose to forego app development and develop a mobile version of their web page often choose to stick with their existing web analytics provider, and oftentimes these providers do not provide a level of detail regarding mobile engagement that would prove particularly useful to marketers who want to capture a snapshot of mobile users. In many cases, companies simply have not given adequate consideration to mobile engagement, despite the fact that it is a growing segment of

online interaction that is only going to grow, particularly as smartphone saturation continues.

However, for those entities that wish to delve further into mobile analytics, there are a growing number of options available, with a few key differences between the major offerings. There are both free and paid mobile analytics platforms available; the key differentiator between these offerings seems to come down to data ownership. A third-party provider that shares the data with you, like Google, is more likely to come at a bargain price, whereas a provider that grants you exclusive ownership of the data is going to come at a premium. Finally, implementation will make a difference in costs: SaaS (software-as-a-service) solutions, which are typically web based, run on the third-party service's own servers, and relatively easy to install, tend to be less expensive, whereas "on-premises" solutions are both rare and quite expensive.

There are a small but growing number of companies that provide app-specific analytic tools, typically deployed as SDKs (software development kits) that can be "hooked" into apps. These companies include, but are by no means limited to, Adobe Analytics, which has been noted for its scalability and depth of analysis; Artisan Mobile, an iOS-focused analytics firm that allows customers to conduct experiments with live users in real time; Bango, which focuses on ad-based monetization of apps; Capptain, which allows specific user segments to be identified and targeted with marketing campaigns; Crittercism, which is positioned as a transaction-monitoring service; Distimo, which aggregates data from a variety of platforms and app stores to create a fuller position of an app in the larger marketplace; ForeSee, which has the ability to record customer interactions with apps; and Kontagent, which touts itself as a tool for maintaining customer retention and loyalty.

As mobile devices and the mobile web grow increasingly sophisticated, there is no doubt that mobile analytics tools will also grow in sophistication. Nevertheless, it would seem that there are a wide range of promising toolkits already available to the marketer who is interested in better understanding customer behaviors and

increasing customer retention, loyalty, and satisfaction.

Cross-References

- [Data Aggregation](#)
- [Network Data](#)

Further Reading

- Meeker, M. *Internet trends 2014*. <http://www.kpcb.com/insights/2014-internet-trends>. Accessed September 2014.
- Smith, A. Smartphone ownership 2013. *Pew Research Center*. <http://www.pewinternet.org/2013/06/05/smartphone-ownership-2013/>. Accessed September 2014.
- Stetler, M. *How much does it cost to develop a mobile app?* AppMuse. <http://appmuse.com/appmusing/how-much-does-it-cost-to-develop-a-mobile-app/>. Accessed September 2014.

Multiprocessing

Joshua Lee
Schar School of Policy and Government, George
Mason University, Fairfax, VA, USA

Synonyms

[Parallel processing](#)

Introduction

Multiprocessing is the utilization of separate processors to complete a given task on a computer. For example, on modern computers, *Microsoft Word* (or just about any executable program) would be a single process. By contrast, multi-threading is done *within* a single process such that a single process can have multiple threads. A multiprocessing approach to computation uses more physical hardware (i.e., additional processors) to improve speed, whereas a multi-threading

approach uses more threads within a single processor to improve speed. Both are meant to optimize performance, but the conditions in which they thrive are different. When utilized correctly, multiprocessing increases overall data throughput whereas multi-threading increases the efficiency and minimizes the idleness of each process.

While the two concepts are closely related, it is important to note that they have significant differences, particularly when it comes to dealing with Big Data. In this entry, we will focus on the core differences that make multiprocessing distinct from multi-threading and what those differences mean for Big Data.

Concurrent Versus Parallel Processing

The terms concurrent and parallel appear frequently in any discussion of multiprocessing, particularly when it is compared with multi-threading. While there is significant overlap between them, they are nevertheless distinct terms. In short, multiple threads are run concurrently, whereas multiple processes are run in parallel. What does this difference actually mean, though?

As an example, imagine that to win a competition you must do 50 push-ups and 50 sit-ups. The concurrent approach to doing this task would be to switch back and forth between them – for example, do one push-up, one sit-up, one push-up, one sit-up, etc. Whereas that would be extraordinarily inefficient for a human being to do, that is what multi-threading does, and it is quite efficient at it. By contrast, the parallel approach to the task would be to have you do all 50 push-ups while having your friend join you in the competition and do 50 sit-ups simultaneously. In this case, the parallel approach would undoubtedly be faster.

Programming for Multiprocessing

In terms of computer programming, multiprocessing offers multiple advantages over multi-threading. First, there are no race conditions with multiprocessing, since each process has a distinct area of memory it operates in (unlike multi-threading, which shares the same memory space). There is no reason to worry about which

process finishes first and thus might modify some bit of memory that another process requires.

Second, it avoids problems with the “global interpreter local” (GIL) utilized by several programming languages such as Python and Ruby. The GIL generally prevents multiple threads from running at once on a single processor when using that programming language. As such, depending on how it is implemented, it can almost negate performance improvements from multi-threading. At the same time, the GIL is used because it can significantly increase the performance of single-threaded performance and allow the programmer not to have to worry about utilizing C-based libraries that are not thread-safe. However, that doesn’t help us when we want to use multi-threading.

Third, child processes are usually easy and safe to interrupt and/or kill when they are done with their task. By contrast, since threads share memory, it can be significantly more complex to safely kill a thread that is done with its task, potentially leaving it wasting resources until the remaining threads are done.

Fourth, there isn’t the issue of deadlock with multiprocessing (as compared to multi-threading). Deadlock occurs when Thread A needs a resource that Thread B has while Thread B needs a resource that Thread A has. When this happens, both threads wait indefinitely for the other thread to drop their resource but never end up doing so. Because of separate memory spaces (and significantly less communication between processes), deadlock doesn’t occur with multiprocessing.

All of this means that, in general, writing efficient code for a multiprocessing environment is much simpler than writing efficient code in a multi-threaded environment. At the same time, the multiprocessing code will require more physical capabilities (i.e., more processing cores) to be able to run.

Performance when Multiprocessing

A multiprocessing approach will often be faster than a multi-threading approach, albeit with some important caveats. This is because there are several features of a given task that can make multiprocessing the slower solution.

First, whereas threads exist in the same memory space and can communicate between one another quite easily, communication between processes requires inter-process communication (IPC), a requirement with vastly greater overhead. If there is a significant requirement for communication between two separate parallel processes, then this additional overhead may slow it down enough for a multi-threading approach to have better performance.

Second, multiple processes require both separate memory spaces and more memory for each process, whereas multiple threads share memory spaces. This can both add additional hardware requirements in terms of memory to run successfully. If the processes themselves are comparatively lightweight, the additional overhead of creating the processes in the first place may outweigh their ability to run faster when comparing performance.

Finally, the required overhead of each process means that having a process lay idle for any significant period becomes a waste of resources. If a given task is only run once during a program's execution, and that run only occurs for 10% of the overall time of the program's execution, it is questionable whether the task should be given its own process by itself due to the wasted resources.

Multiprocessing Versus Multi-Threading Usage with Big Data

With a review of programming and performance issues in hand, what does this mean when dealing with Big Data? While there are endless kinds of programming tasks that might need to be done, in general, we can abstract these tasks out to either I/O tasks or raw computational tasks. An I/O task is one where the inputting and outputting of data is central to completing the task. This would include frequently writing to or reading from files, scraping large quantities of web pages of the internet, or accepting user input from the keyboard, among other tasks. By contrast, a more computational task in this context could involve sorting/searching through massive troves of data, performing machine learning training or inference, or applying any kind of *en masse* mathematical transformation to one's data. While there may

be tasks that don't easily fit into either one, this basic typology should provide a starting point for when to use multiprocessing.

In general, multi-threading should be utilized for I/O tasks, whereas multiprocessing should be utilized for computational tasks. This split mainly derives from the issues of overhead, idleness, and IPC. Multiprocessing requires significantly more overhead, which means that I/O tasks being idle and/or requiring communication between separate processes during computation would be a significant slowdown. If your processes are lying idle for any significant stretch of time, it is simply a waste of resources. When it comes to I/O tasks, these idle times can be significantly more common because input and output (usually) are not constantly occurring in the way that, say, training a machine learning model is almost never idle.

Consider the case of a web page: a web server can have multiple tasks going on simultaneously. To name a few, this includes tracking user clicks, sending information to back-end databases, and displaying the proper HTML in the users' browser. However, none of these individual tasks are being run constantly – the user is not continually making clicks, back-end databases don't need to be constantly appended every millisecond, and new HTML doesn't need to be constantly displayed to the user. Were we to assign three separate processes for each of these three tasks, many of those processes would remain idle for significant amounts of time, wasting valuable computing resources. By contrast, a concurrent multi-threaded approach would be much more fitting on a web server because all the tasks on a server don't need to be executed all the time, even if one or more threads is idle, the others in the process can take up the slack and make most efficient use of a single process.

Conclusion

In conclusion, multi-threading and multiprocessing both have their place at the Big Data table. Neither is a perfect solution for all occasions, but each has circumstances in which it is superior. While using multiprocessing at the wrong time can lead to massive wastes of computational resources, using multi-threading at the wrong

time can lead to no improvement (or even decreases) in performance. Indeed, the purposeful and careful combination of multi-threading together with multiprocessing (that is, multiple processes each with multiple threads) will ensure optimal performance for a wide array of Big Data-oriented tasks.

Cross-References

- ▶ [Multi-Threading](#)
- ▶ [Parallel Processing](#)

Further Reading

- Bellairs, R. (2019, April 10). *How to take advantage of multithreaded programming and parallel programming in C/C++*. Retrieved from PERFORCE: [https://www.perforce.com/blog/qac/multithreading-parallel-programming-c-cpp#:~:text=Parallel%20programming%20is%20a%20broad,set%20\(thread\)%20of%20instructions.&text=These%20threads%20could%20run%20on%20a%20single%20processor](https://www.perforce.com/blog/qac/multithreading-parallel-programming-c-cpp#:~:text=Parallel%20programming%20is%20a%20broad,set%20(thread)%20of%20instructions.&text=These%20threads%20could%20run%20on%20a%20single%20processor).
- Nagarajan, M. (2019, December 2). *Concurrency vs. parallelism — A brief view*. Retrieved from Medium: <https://medium.com/@itIsMadhavan/concurrency-vs-parallelism-a-brief-review-b337c8dac350>.
- Rodrigues, G. S. (2020, September 27). *Multithreading vs. multiprocessing in Python*. Retrieved from Towards Data Science: <https://towardsdatascience.com/multithreading-vs-multiprocessing-in-python-3afeb73e105f>.

Multi-threading

Joshua Lee
Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Introduction

Multi-threading is the utilization of multiple *threads* to complete a given task (i.e., process) in parallel. It is one of the fundamental mechanisms through which data processing and execution can be substantially sped up. However, utilizing multi-threading properly is more complicated than

simply hitting an “on” switch. In fact, if it is poorly implemented, multi-threading may produce no discernable impact or even *decrease* the performance of certain tasks (<https://brooker.co.za/blog/2014/12/06/random.html>). Therefore, any project working with Big Data should thoroughly study multi-threading before implementing it.

A Basic Example

Let **process A** need to divide 500 integers by 5 and print the results for each. Without multi-threading (i.e., with a single thread), the integers are divided serially (i.e., one at a time) until all 500 operations are complete. By contrast, with multi-threading, the integers are divided in parallel by different threads. If the process uses 2 threads, it would split the 500 integers into 2 lists of 250 each. Then, each thread would work on their assigned list. Theoretically, this could double the completion speed. However, realistically this speedup will be lower due to the additional processing overhead from the additional thread and if there are sections of code that can’t be parallelized.

Conceptual Foundations

Understanding multi-threading requires an understanding of several other computer science concepts.

Process – A computer program that is currently running. For example, *Microsoft Excel* becomes a process when a user double-clicks on its icon.

Thread – A series of instructions within a process that can be executed independently of any other code in that process.

Thread-safe – A code which is thread-safe is implemented in such a way as to allow for multiple threads to access and modify its data safely, such as to maintain data integrity. While all codes would be thread-safe in a perfect world, there are practical limitations that prevent this (that discussion is beyond the scope of this section).

Lightweight Process (LWP) – Generally used as an alternate name for a thread. This is because a thread is essentially a process (they can accomplish the same tasks in theory), but a thread is a

more restricted process that has less overhead and can't run processes inside it.

Overhead – Indirect or excess computation time. Excessive overhead will substantially slow down the performance. It also increases linearly – four threads will have roughly four times the overhead as one thread.

Process Versus Thread

While theoretically, one process and one thread could be instructed to perform the same task, they have some key distinctions that affect how they're utilized. Specifically:

- Processes are generally used for “heavy-weight,” major tasks, whereas threads are generally used for “lightweight,” minor tasks. This is because a process can have one or more threads inside it, but a thread cannot have a process inside.
- A process has much larger *overhead* than a thread – starting up and managing a new process is itself a computationally intensive task that can slow down the performance.
- Different threads within a single process share the same *address space* (memory), whereas different processes on an operating system do not. Sharing the same address space allows different threads to access the same variables in memory and to communicate with one another quickly and easily. By contrast, sharing information between processes (known as inter-process communication, or IPC) is far more computationally intensive.
- However, using multiple processes (versus multiple threads) allows for more isolation for each process – processes cannot directly interact with each other's memory/variables, which can be useful for some tasks. This adds an inherent layer of security between processes that don't exist between threads.

Core Versus Thread

One potentially confusing aspect of understanding multi-threading is the relationship between threads and cores. A normal single core can run only a single thread at a time. However, even on a single-core machine, by swiftly moving back and

forth between different processes (each of which has at least one thread), the machine creates the illusion to the user that all processes are being run simultaneously.

Because of this, it is sometimes possible to efficiently use more threads than there are cores. For example, assume there is a dual core machine with core A and core B. In addition, there is thread A, B, C, and D; all four of which are contained within process A. Threads A and B are running on core A, and threads C and D are running on core B. If threads A and C are frequently sleeping or waiting, then their respective cores can run threads B and D during this period, allowing for the efficient use of more threads than there are cores.

Threads and Shared Memory Access

One of the greatest benefits of multi-threading is that the threads share memory access and can thus work together to complete the same task. However, this benefit also has a substantial drawback that must be taken into consideration. What happens if two threads running in a single process determine that they need to modify the same addressed memory (i.e., variable) at the same time? What determines which thread should get priority, and how is this conflict managed? Solutions to this problem are classified under the term *thread synchronization*.

Different programming languages utilize different solutions to the issue of shared memory access. One of the most common solutions is via mutual exclusion (mutex). With mutex, an object (i.e., memory/variable/address space) is “locked” by one thread. Any other thread which attempts to access the locked object is refused access. Other methods of synchronization, beyond the scope of this guide, include barriers, semaphores, and spinlocks ([https://msdn.microsoft.com/en-us/library/ms228964\(v=vs.110\).aspx](https://msdn.microsoft.com/en-us/library/ms228964(v=vs.110).aspx)).

There are many varieties of mutual exclusion, but two of the most common are *queuing mutex* and *read/write mutex*, also known as *shared mutex*. A queuing mutex creates a FIFO framework for threads requesting a locked object. For

example, let thread A have the lock on object Z, and let thread B request access to object Z. Under queuing mutex, thread B would “wait in line” for object Z to become available. Then, if thread C also requested access to object Z, it would need to get in line behind thread B. While this has the advantage of the simplicity of understanding, it can also create additional overhead for the operating system to manage, as well as threads excessively waiting in line rather than processing information.

A read/write mutex, also known as a *shared mutex*, allows for any number of threads to *read* an object simultaneously. However, if a thread wants to *write* to that object, it must wait until all the threads currently reading it have let go of their locks.

Common Multi-threading Design Patterns

There are many common design patterns for multi-threaded programming, which are naturally efficient. Generally, pre-existing design patterns should be considered before attempting to invent a new one.

Boss-Worker Thread Pattern: In the boss-worker thread pattern, there is one thread, which is the “boss,” and all other threads are “workers.” When new tasks need to be completed, the boss thread assigns the task to a given worker thread, creating a new thread on the spot if none are available. This is one of the simplest and most common patterns – it allows for ease of use and ease of debugging. However, it can also create problems of contention between threads if they require interdependent resources.

Pipeline Pattern: In the pipeline pattern, each thread completes a portion of a given task and then passes it on to the next thread. This is also a simple pattern and can be most useful when there are discrete steps that need to be completed that are sequential in nature. However, it can also require substantial fine tuning to ensure that each stage of the pipeline doesn’t cause a bottleneck. Additionally, the parallelization that can occur is limited by the number of pipelines.

Common Pitfalls in Multi-threading

Implementing a multi-threaded design also comes with certain pitfalls to avoid. Different types of mutex locking and different design patterns must deal with these pitfalls to varying degrees, but they always need to be taken into design consideration.

Race Conditions: A race condition is where thread B’s processing interferes with thread A’s processing due to them both being run simultaneously. For example, consider functions X and Y below:

```
Function X(integer C){
    if C == 5:
        return True;
    else:
        return False;
}
Function Y(integer C){
    C = C + 1;
    return C;
}
```

Next, let thread A call X, thread B call Y, and integer C start at 5. If the threads are set to run their functions at the same time, will X return true or false? The answer is, it depends – it’s an unstable race between the two threads.

Thus, if we run this experiment 100 times, there will not be consistency in the result. Threads A and B are racing against one another to decide the result. If thread A happens to finish processing fast enough on one run, it will return true. But sometimes, thread B will run fast enough that thread A will return false. This kind of inconsistent result, given the same conditions and starting point, can cause difficult-to-spot bugs in a multi-threaded program.

Deadlocks: Deadlocks occur when no single thread can execute. Let thread A have a lock for object Z and thread B have the lock for object Y. Thread A will only give up object Z when it can grab the lock for object Y, and thread B will only give up object Y when it can grab the lock for object Z. In this situation, both threads will wait for eternity because neither of their conditions can be fulfilled.

Multi-threaded Design Optimization

Even after shared memory access issues are resolved, pitfalls are compensated for, and a design pattern is chosen, performance can still be further optimized. Especially when dealing with Big Data, even a minor performance increase can substantially impact processing time. Below are some of the most important optimizations to consider:

Granularity: Granularity is a measurement for how much *real work* is done in each thread. Threads that are sleeping or waiting are not performing real work. For example, we need a program to square 800 integers. *Fine granularity* would be if more threads each accomplished less work. Thus, the maximum fine granularity would be if 800 threads each performed one squaring operation. Needless to say, this isn't an efficient design. By contrast, *coarse granularity* would be if few threads each accomplished more work. Thus, maximum coarse granularity would be not to use multi-threading at all.

If the threads are too fine, it creates unnecessary overhead from handling the threads themselves. However, if granularity is too coarse, threads can suffer from a load imbalance – for example, one thread can take 1 h to complete its tasks, whereas another thread only takes 10 min. In that case, the application itself would still take an hour to complete, even though one thread is sitting idly by for most of that period. Therefore, granularity optimization aims to find the proper balance between the two extremes, both in terms of load balancing and overhead minimization.

Lock Ordering: One method of avoiding *dead-lock* is lock ordering. With lock ordering, locks should be obtained in a fixed order throughout the program. This order is determined by what other

threads will need those locks and when they will need it.

Lock Frequency: The act of locking and unlocking itself adds overhead. Analyze your program to see if there are perhaps ways you can minimize this frequency.

Critical Sections: A critical section is a part of the code which must be accomplished serially (i.e., in order and without multiple threads). These sections are naturally time-consuming in any multi-threaded algorithm. Minimizing the size and computational complexity of these critical sections is vital for optimizing performance.

Conclusion

The most important point to remember about utilizing a multi-threaded design is that *it's not the solution for every problem*. Furthermore, there are numerous structural factors that can inhibit its effectiveness, and the decision on whether to utilize a multi-threaded design should be handled with care. Even after the decision to use a multi-threaded design is made, it may require substantial optimization to obtain the desired performance enhancements.

Further Reading

- Lewis, B., & Berg, D. J. (1995). *Threads primer: A guide to multithreaded programming*. Upper Saddle River, NJ: Prentice Hall Press.
- Protopopov, B. V. (1996). *Concurrency, multi-threading, and message passing*. Master's thesis, Department of Computer Science, Mississippi State University.
- Ungerer, T., Robič, B., & Šilc, J. (2003). A survey of processors with explicit multithreading. *ACM Computing Surveys (CSUR)*, 35(1), 29–63.

N

National Association for the Advancement of Colored People

Steven J. Campbell
University of South Carolina, Lancaster,
Lancaster, SC, USA

The National Association for the Advancement of Colored People (NAACP) is an African-American civil rights organization headquartered in Baltimore, MD. Founded in 1909, its membership advocates civil rights by engaging in activities such as mobilizing voters and tracking equal opportunity in government, industry, and communities. Over the past few years, the NAACP has shifted its attention to digital advocacy and the utilization of datasets to better mobilize activists online. In the process, the NAACP has become a leading organization in how it harnesses big data for digital advocacy and related campaigns. The NAACP's application of specially tailored data to its digital approach, from rapid response to targeted messaging to understanding recipients' interests, has become an example for other groups to follow. At the same time, the NAACP has challenged other big data (both in the public and private sectors), highlighting abuse of such data in

ways that can directly impact disadvantaged minority groups.

With a membership of over 425,000 members, the NAACP is the nation's largest civil rights organization. Administered by a 64-member board headed by a chairperson, various departments within the NAACP govern particular areas of action. The Legal Department tracks court cases with potentially extensive implications for minorities, including recurring discrimination in areas such as education and employment. The Washington, D.C., office lobbies Congress and the Presidency on a wide range of policies and issues, while the Education Department seeks improvements in the sphere of public education. Overall, the NAACP's mission is to bolster equal rights for all people in political, educational, and economic terms as well as stamp out racial biases and discrimination.

In order to extend this mission into the twenty-first century, the NAACP launched a digital media department in 2011. This entailed a mobile subscriber project that led to 423,000 contacts, 233,000 Facebook supporters, and 1.3 million email subscribers, due in large part to greater social media outreach. The NAACP's "This is my Vote!" campaign, launched prior to the 2012 presidential election, dramatically advanced the organization's voter registration and mobilization

programs. As a result, the NAACP registered twice the number of individuals – over 374,000 – than it did in 2008 and mobilized over 1.2 million voters. In addition, the NAACP conducted an election eve poll that surveyed 1,600 African-American voters. This was done in order to assess their potential influence as well as key issue areas prior to the election results and in looking forward to 2016. Data from the poll highlighted the predominate role played by African-Americans in major battleground states and divulged openings for the Republican Party in building rapport with the African-American community. In addition, the data signaled to Democrats a message not to assume levels of Black support in 2016 on par with that realized in the 2008 and 2012 elections.

By tailoring its outreach to individuals, the NAACP has been successful in achieving relatively high rates of engagement. The organization segments supporters based on their actions, such as whether they support a particular issue based on past involvement. For instance, many NAACP members view gun violence as a serious problem in today's society. If such a member connects with NAACP's online community via a particular webpage or internet advertisement, s/he will be recognized as one espousing stronger gun control laws. Future outreach will entail tailored messages expressing attributes that resonate on a personal level with the supporter, not unlike that from a friend or colleague.

The NAACP also takes advantage of major events that reflect aspects of the organization's mission statement. Preparation for such moments entails much advance work, as evidenced in the George Zimmerman trial involving the fatal shooting of 17-year-old Trayvon Martin. As the trial was concluding in 2013, the NAACP formed contingency plans in advance of the court's decision. Website landing pages and prewritten emails were set in place, adapted for whatever result may come. Once the verdict was read, the NAACP sent out emails within 5 min that detailed specific actions for supporters to take. This resulted in over a million petition signatures demanding action on the part of the US Justice Department, which it eventually took.

Controversy

While government and commercial surveillance potentially affect all Americans, minorities face these risks at disproportionate rates. Thus, the NAACP has raised concerns about whether big data needs to provide greater protections for minorities in addition to the general privacy protections commonly granted. Such controversy surrounding civil rights and big data may not be self-evident; however, big data often involves the targeting and segmenting of one type of individual from another. This serves as a threat to basic civil rights – which are protected by law – in ways that were inconceivable in recent decades. For instance, the NAACP has expressed alarm regarding the collection of information by credit reporting agencies. Such collections can result in the making of demographic profiles and stereotypical categories, leading to the marketing of predatory financial instruments to minority groups.

The US government's collection of massive phone records for purposes of intelligence has also drawn harsh criticism from the NAACP as well as other civil rights organizations. They have vented warnings regarding such big data by highlighting how abuses can uniquely affect disadvantaged minorities. The NAACP supports principles aimed at curtailing the pervasive use of data in areas such as law enforcement and employment. Increasing collections of data are viewed by the NAACP as a threat since such big data could allow for unjust targeting of, and discrimination against, African-Americans. Thus, the NAACP strongly advocates measures such as a stop to "high-tech profiling," greater pressure on private industry for more open and transparent data, and greater protections for individuals from inaccurate data.

Cross-References

- [Demographic Data](#)
- [Facebook](#)
- [Pattern Recognition](#)

Further Reading

Fung, Brian (27 Feb 2014). Why civil rights groups are warning against 'big data'. *Washington Post*. <http://www.washingtonpost.com/blogs/the-switch/wp/2014/02/27/why-civil-rights-groups-are-warning-against-big-data/>. Accessed Sept 2014.

Murray, Ben (3 Dec 2013). What brands can learn about data from the NAACP: Some advocacy groups are ahead of the curve, making smarter data decisions. *Advertising Age*. <http://adage.com/article/datadriven-marketing/brands-learn-data-advocacy-groups/245498/>. Accessed Sept 2014).

NAACP. <http://www.NAACP.org>. Accessed Sept 2014.

National Oceanic and Atmospheric Administration

Steven J. Campbell
University of South Carolina Lancaster,
Lancaster, SC, USA

The National Oceanic and Atmospheric Administration (NOAA) is an agency housed within the US Commerce Department that monitors the status and conditions of the oceans and the atmosphere. NOAA oversees a diverse array of satellites, buoys, ships, aircraft, tide gauges, and supercomputers in order to closely track environmental changes and conditions. This network yields valuable and critical data that is crucial for alerting the public to potential harm and protecting the environment nationwide. The vast sums of data collected daily have served as a challenge to NOAA in storing as well as making the information readily accessible and meaningful to the public and interested organizations. In the future, as demand grows for ever-greater amounts and types of climate data, NOAA must be resourceful in meeting the demands of public officials and other interested parties.

First proposed by President Richard Nixon, who wanted a new department in order to better protect citizens and their property from natural dangers, NOAA was founded in October 1970. Its mission is to comprehend and foresee variations in the environment, from the conditions of

the oceans to the state of the sun, and to better safeguard and preserve seashores and marine life. NOAA provides alerts to dangerous weather, maps the oceans and atmosphere, and directs the responsible handling and safeguarding of the seas and coastal assets. One key way NOAA pursues its mission is by conducting research in order to further awareness and better management of environmental resources. With a workforce of over 12,000, NOAA consists of six major line offices, including the National Weather Service (NWS), in addition to over a dozen staff offices.

NOAA's collection and dissemination of vast sums of data on the climate and environment contribute to a multibillion-dollar weather enterprise in the private sector. The agency has sought ways to release extensive new troves of this data, an effort that could be of great service to industry and those engaged in research. NOAA announced a call in early 2014 for ideas from the private sector to assist the agency's efforts in freeing up a large amount of the 20 terabytes of data that it collects on a daily basis pertaining to the environment and climate change. In exchange, researchers stand to gain critical access to important information about the planet, and private companies can receive help and assistance in advancing new climate tools and assessments.

This request by NOAA shows that it is planning to place large amounts of its data into the cloud, benefitting both the private and public sectors in a number of ways. For instance, climate data collected by NOAA is currently employed for forecasting the weather over a week in advance. In addition, marine navigation and offshore oil and gas drilling operations are very interested in related data. NOAA has pursued unleashing ever-greater amounts of its ocean and atmospheric data by partnering with groups outside government. This is seen as paramount to NOAA's data management, where tens of petabytes of information are recorded in various ways, engendering over 15 million results daily – from weather forecasts for US cities to coastal tide monitoring – which totals twice the amount of all the printed collections of the US Library of Congress.

Maneuvering through NOAA's mountain of weather and climate, the data has proved to be a great challenge over the years. To help address this issue, NOAA made available, in late 2013, an instrument that helped further open up the data to the public. With a few clicks of a mouse, individuals can create interactive maps illustrating natural and manmade changes in the environment worldwide. For the most part, the data is free to the public, but much of the information has not always been organized in a user-friendly format. NOAA's objective was to bypass that issue and allow public exploration of environmental conditions from hurricane occurrences to coastal tides to cloud formations. The new instrument, named NOAA View, allows ready access to many of NOAA's databases, including simulations of future climate models. These datasets grant users the ability to browse various maps and information by subject and time frame. Behind the scenes, numerous computer programs manipulate datasets into maps that can demonstrate environmental attributes and climate change over time. NOAA View's origins were rooted in data visualization instruments present on the web, and it is operational on tablets and smartphones that account for 44% of all hours spent online by the US public.

Advances to NOAA's National Weather Service supercomputers have allowed for much faster calculations of complex computer models, resulting in more accurate weather forecasts. The ability of these enhanced supercomputers to analyze mounds of scientific data proves vital in helping public officials, communities, and industrial groups to better comprehend and prepare for perils linked with turbulent weather and climatic occurrences. Located in Virginia, the supercomputers operate with 213 teraflops (TF) – up from the 90 TF with the computers that came before them. This has helped to produce an advanced Hurricane Weather Research and Forecasting (HWRF) model that the National Weather Service can more effectively employ. By allowing more effective monitoring of violent storms and more accurate predictions regarding the time, place, and intensity of their impact, the HWRF model can result in saved lives.

NOAA's efforts to build a Weather-Ready Nation have evolved from a foundation of super-computer advancements that have permitted more accurate storm-tracking algorithms for weather prediction. First launched in 2011, this initiative on the part of NOAA has resulted in advanced services, particularly in ways that data and information can be made available to the public, government agencies, and private industry.

Cross-References

- ▶ [Climate Change, Hurricanes/Typhoons/Cyclones](#)
- ▶ [Cloud Computing](#)
- ▶ [Data Storage](#)
- ▶ [Environment](#)
- ▶ [Predictive Analytics](#)

Further Reading

- Freedman, A. (2014, February 24). *U.S. readies big-data dump on climate and weather*. <http://mashable.com/2014/02/24/NOAA-data-cloud/>. Accessed September 2014.
- Kahn, B. (2013). *NOAA's new cool tool puts climate on view for all*. <http://www.climatecentral.org/news/noaas-new-cool-tool-puts-climate-on-view-for-all-16703>. Accessed September 2014.
- National Oceanic and Atmospheric Administration (NOAA). www.noaa.gov. Accessed September 2014.

National Organization for Women

Deborah Elizabeth Cohen
Smithsonian Center for Learning and Digital
Access, Washington, DC, USA

The National Organization for Women (NOW) is an American feminist organization that is the grassroots arm of the women's movement and the largest organization of feminist activists in the United States. Since its founding in 1966, NOW has engaged in activity to bring about

equality for all women. NOW has been participating in recent dialogues to identify how common big data working methods lead to discriminatory practices against protected classes including women. This entry discusses NOW's mission and issues related to big data and the activities NOW has been involved with to end discriminatory practices resulting from the usage of big data.

As written in its original statement of purpose, the purpose of NOW is to take action to bring women into full participation in the mainstream of American society, exercising privileges and responsibilities in completely equal partnership with men. NOW strives to make change through a number of activities including lobbying, rallies, marches, and conferences. NOW's six core issues are economic justice, promoting diversity and ending racism, lesbian rights, ending violence against women, constitutional equality, and access to abortion and reproductive health.

NOW's current president Terry O'Neill has stated that big data practices can render obsolete the USA's landmark civil rights and anti-discrimination laws with special challenges for women, the poor, people of color, trans-people, and the LGBT community. While the technologies of automated decision-making are hidden and largely not understood by average people, they are being conducted with an increasing level of pervasiveness and used in contexts that affect individuals' access to health, education, employment, credit, and products. Problems with big data practices include the following:

- Big data technology is increasingly being used to assign people to ideologically or culturally segregated clusters, profiling them and in doing so leaving room for discrimination.
- Through the practice of data fusion, big data tools can reveal intimate personal details, eroding personal privacy.
- As people are often unaware of this "scoring" activity, it can be hard for individuals to break out of being mislabeled.
- Employment decisions made through data mining have the potential to be discriminatory.
- Metadata collection renders legal protection of civil rights and liberties less enforceable, undoing civil rights law.

Comprehensive US civil rights legislation in the 1960s and 1970s resulted from social actions organized to combat discrimination. A number of current big data practices are in misalignment with these laws and can lead to discriminatory outcomes.

NOW has been involved with several important actions in response to these recognized problems with big data. In January of 2014, the US White House engaged in a 90-day review of big data and privacy issues, to which NOW as a participating stakeholder provided input. Numerous policy recommendations resulted from this process especially related to data privacy and the need for the federal government to develop technical expertise to stop discrimination.

The NOW Foundation also belongs to a coalition of 200 progressive organizations named the Leadership Conference on Civil and Human Rights whose mission is to promote the civil and human right of all persons in the United States. NOW President Terry O'Neill serves on the Coalition's Board of Directors. In February 2014, The Leadership Conference released five "Civil Rights Principles for the Era of Big Data" and in August 2014 provided testimony based on their work to the US National Telecommunications and Information Administration's Request for Public Comment related to Big Data and Consumer Privacy. The five civil rights principles to ensure that big data is designed and used in ways that respect the values of equal opportunity and equal justice include the following:

1. Stop high tech profiling – ensure that clear limits and audit mechanisms are in place to make sure that data gathering and surveillance tools that can assemble detailed information about a person or group are used in a responsible and fair way.
2. Ensure fairness in automated decisions – require through independent review and

other measures that computerized decision-making systems in areas such as employment, health, education, and lending operate fairly for all people and protect the interests of those that are disadvantaged and have historically been discriminated against. Systems that are blind to preexisting disparities can easily reach decisions that reinforce existing inequities.

3. Preserve constitutional principles – government databases must not be allowed to undermine core legal protections, including those of privacy and freedom of association. Independent oversight of law enforcement is particularly important for minorities who often receive disproportionate scrutiny.
4. Enhance individual control of personal information – individuals, and in particular those in vulnerable populations including women and the LGBT community, should have meaningful and flexible control over how a corporation gathers data from them and how it uses and shares that data. Nonpublic information should not be shared with the government without judicial process.
5. Protect people from inaccurate data – Government and corporate databases must allow everyone to appropriately ensure the accuracy of personal information used to make important decisions about them. This requires disclosure of the data and the right to correct it when inaccurate.

Big data has been called the civil rights battle of our time. Consistent with its mission, NOW is engaged in this battle, protecting civil rights of women and others against discriminatory practices that can result from current big data practices.

Cross-References

- [Data Fusion](#)
- [Data Mining](#)
- [National Oceanic and Atmospheric Administration](#)
- [White House Big Data Initiative](#)

Further Reading

- Big data: Seizing opportunities, preserving values.* (2014). Washington, DC: The White House. www.whitehouse.gov/sites/default/files/docs/big-data-privacy-report-5.1.1.14-final-print.pdf. Accessed 7 Sep 2014.
- Eubanks, V. (2014). How big data could undo our civil rights laws. *The American Prospect*. www.prospect.org/article/how-big-data-could-undo-our-civil-rights-laws. Accessed 7 Sep 2014.
- Gangadharan, S. P. (2014). The dangers of high-tech profiling, using big data. *The New York Times*. www.nytimes.com/roomfordebate/2014/08/06/Is-big-data-spreading-inequality/the-dangers-of-high-tech-profiling-using-big-data. Accessed 5 Sep 2014.
- NOW website. (2014). *Who we are*. National Organization for Women. <http://now.org/about/who-we-are/>. Accessed 2 Sep 2014.
- The Leadership Conference on Civil and Human Rights. (2014). *Civil rights principles for the era of big data*. www.civilrights.org/press/2014/civil-rights-principles-big-data.html. Accessed 7 Sep 2014.

National Security Administration (NSA)

► [Data Mining](#)

National Security Agency (NSA)

Doug Tewksbury
Communication Studies Department, Niagara
University, Niagara, NY, USA

The National Security Agency (NSA) is the US governmental agency responsible for collecting, processing, analyzing, and distributing signal-based intelligence information to support military and national security operations, as well as providing information security for US governmental agencies and its allies. Alongside the Central Security Service (CSS), which serves as a liaison between the NSA and military intelligence-gathering agencies, the NSA/CSS serves as 1 of 17 intelligence agencies in the American government, reporting equally to the Department of

Defense and the Director of National Intelligence. Its central mission is to use information gathered through surveillance and codebreaking to support the interests of the United States and its allies.

The NSA has become the center of a larger debate over the proper extent of state surveillance powers in balancing both national security and civil liberties. As the world has become increasingly globalized, and as cultural expression has increasingly become mediated through information flows and new technological developments, the NSA has seen its importance in the national intelligence-gathering landscape rise in tandem with its ability to collect, store, and analyze information through mass surveillance of electronic communications.

This tension became particularly fervent following former NSA contractor and whistleblower Edward Snowden's 2013 revelation that the agency had been secretly collecting the internet, telephone, mobile location, and other digital records of over a billion people worldwide, including tens of millions of domestically based US citizens and dozens of heads of state of foreign governments. Many of the NSA's surveillance practices require no court approval, oversight, or warrant issuing: There is considerable legal disagreement on whether these warrantless collections violate Fourth Amendment protections against search and seizure. The secret Foreign Intelligence Surveillance Court (FISC) that oversees many of the NSA's data-collection strategies has repeatedly allowed these practices. However, the rulings from FISC courts are classified, neither available to the public or most members of Congress, and there have been contradictory rulings from lower and appeals courts on the FISC's interpretation of law. The US Supreme Court is expected to address these issues in the near future, but as of this writing, it has not yet ruled on the constitutionality of most of the NSA's surveillance practices. Most of what is known about the NSA's activities has thus far come from the Snowden leaks and subsequent interpretation of the leaked documents by media organizations and the public. However, the full extent of the NSA's practices continues to be unknown.

Agency History and Operations

The National Security Agency was created in 1952, evolving out of the Cipher Bureau and Military Intelligence Branch, a World War I-era cryptanalytic agency, and later, the Armed Forces Security Agency, both of which dealt with the encryption of the messages of American forces and its allies through the end of the Second World War. The mandate of the organization continues to be one of signal intelligence – mediated, signal-based information sources such as textual, radio, broadcast, or telephonic communications – rather than human intelligence, which is the domain of the Central Intelligence Agency (CIA) and other governmental agencies. Though the NSA's existence was classified upon the agency's creation, and its practices clandestine, it would become controversial in the 1960s and 1970s for its role in providing evidence for the Gulf of Tonkin incident, domestic wiretaps of anti-Vietnam War protesters and civil rights leaders, the agency's involvement with the Watergate scandal of the Nixon Administration, and numerous military actions of the United States and economic espionage instances during the 1980s and 1990s. Both the NSA's budget and number of employees are classified information, but in 2016 were estimated to be just under \$10b and between 35,000 and 45,000, respectively. Its headquarters is in Fort Meade, Maryland.

The NSA in the Twenty-First Century

Technological Capabilities

The per-bit cost of storage continues to decrease dramatically with each passing year while processing speed increases exponentially. With access to both the deep pockets of the US Government and the data infrastructure of American ISPs, the technological and logistical capabilities of the NSA continue to lead to new programs of surveillance and countersurveillance, often at the leading edge of technological and scientific discovery.

In terms of its global advantages in these terms of data collection and processing, what is known

about the NSA reads as a list of superlatives: It has more combined computing power, more data storage, the largest collection of supercomputers, and more taps on global telephone and internet connections than any other governmental or private entity in the world. Particularly following the 2013 opening of its 1 million square foot Utah Data Center outside of Salt Lake City, potentially holding upward of a yottabyte of data, it is estimated that the NSA now has the ability to surveil most of the world's internet traffic, most notably through the signals that run through public and private servers in the United States. The NSA has numerous facilities throughout the United States, around the globe in allied nations, and at least four spy satellites dedicated for its exclusive use. It has spent at least hundreds of millions of dollars to fund the development of quantum computing platforms that, if realized, will be able to decrypt the most complex algorithmic encryption available today. Billions of the world's emails, computer data transfers, text messages, faxes, and phone calls flow through the NSA's computing centers every hour, many of which are logged and indexed.

Surveillance and Countersurveillance Activities

In June 2013, *The Guardian* reported that they had received documents leaked by former NSA contractor Edward Snowden that detailed that the FISC had secretly ordered Verizon Communications to provide the NSA a daily report for all calls made in its system by its 120 million customers, both within the United States and between the United States and other countries and, in bulk, with no discrimination based on suspicion of wrongdoing. While the content of the calls was not included, the corporation handed over the call's metadata: the numbers involved, geographic location data, duration, time, routing information, and other transactional data. These practices had existed in some form for over a decade under the Bush and Obama Administrations through the also-controversial "warrantless wiretapping" provisions of the USA PATRIOT Act. But in this case, many in Congress and the public were

surprised at the extent of the NSA's data collection and retention, as the organization is prevented from knowingly surveilling US citizens on US soil. However, the mass collection of data has often been indiscriminate, and an unknown number of unintentional targets were regularly swept up in the collection. In March 2014, President Obama announced slight alterations to the NSA's bulk telephone metadata collection practices, but these did little to quell the controversy or appease the public, a majority of whom continued to oppose the agency's domestic surveillance practices as recently as 2016.

Beyond the telephone metadata collection, the NSA's data-collection and analysis activities are numerous and include such programs as PRISM, MUSCULAR, Boundless Informant, XKEYSCORE, and several known others. These have produced similar massive databases of user information for both foreign and non-foreign users and often with the collaboration between the NSA and other foreign (primarily European) intelligence agencies. It has been documented that a large number of US service providers have given the NSA information directly from their servers or through direct access to their network lines, including Microsoft, Yahoo, Google, Facebook, PalTalk, AOL, Skype, YouTube, Apple, and AT&T.

The MYSTIC program collected metadata from a number of nation-states' territories, apparently without the consent of the governments, and used in-house developed voice-recognition software under the subsequent SOMALGET program to record both full-take audio and metadata for every telephone conversation in Bermuda, Iraq, Syria, and others. The NSA also intentionally weakened the security of a number of encryption protocols or influenced the production of a master encryption key in order to maintain a "back door" through its BULLRUN program.

The NSA regularly intercepts server and routing hardware – most of which is built by US corporations – after they are shipped via postal mail, but before they are delivered to government or private recipients in countries, implants hardware or software surveillance tools and then repackages them with a factory seal and sends

them onward, allowing post-encryption access to the information sent through them.

Edward Snowden revealed in 2014 that the NSA also routinely hacks foreign nations' networks, not only military or governmental servers but also academic, industrial, corporate, or medical facilities. NSA hackers, for example, attempting to gain access to one of the core routers in a Syrian ISP in 2012, crashed the ISP's routing system, which in turn cascaded and blacked out the entire nation's internet access for several days.

The SEXINT program has been monitoring and indexing the sexual preferences and pornography habits of internet users, political activists, and dissidents in order to "call into question a radicalizer's dedication" to a cause by releasing the potentially embarrassing details.

The NSA has admitted that it monitored the personal cell phones and electronic communication of at least 35 world leaders (including many nations allied with the United States), as well as attendees to the 2010 G20 Conference in Toronto, EU embassies in Washington, DC, visiting foreign diplomats, and apparently many others, all without their knowledge. It has collected massive indiscriminate datasets of foreign citizens' communications, including 45 million Italian phone calls, 500 million German communications, 60 million Spanish phone calls, 70 million French phone calls, 33 million Norwegian communications, and hundreds of millions of Brazilian communications in 30-day increments in 2012 and 2013.

Furthermore, it was revealed in mid-2014 that the NSA had implemented its AI platform MonsterMind, which is designed to detect cyber attacks, block them from entering the U.S., and automatically counterattack with no human involvement, a problematic practice that, according to Snowden, requires the interception of all traffic flows in order to analyze threats.

Legal Oversight

There have been questions over the legality of many of the National Security Agency's practices, particularly in terms of the possibility of civil rights abuses that can occur without adequate public transparency and oversight, both

domestically in the United States and worldwide. Most of the agency's data-collection practices are clandestine and fall under the jurisdiction of the Federal Intelligence Surveillance Court, a secret, non-adversarial court that rules on the constitutionality of US governmental agencies' surveillance practices. The FISC has, itself, been critiqued for its secrecy and lack of transparency and accountability, both from members of the public and from Congress, as well as its critique as a "rubber stamp" court that approves nearly all of the requests that the government submits.

US citizens have constitutional protections that are not granted to noncitizens, and many within the country have argued that the mass surveillance of Americans' telephone, internet, and other activities is a violation of the Fourth Amendment's prohibition against illegal search and seizure. Others have upheld the authority of the FISC's rulings and need for secrecy in the name of national security, particularly in an age where violent and cyber terrorism are prescient threats. The NSA requires that its intelligence analysts have 51% confidence in their target's "foreignness" for data collection, and many American citizens are routinely swept up in massive intelligence gathering. It was reported in 2013 that the agency shares its raw data with the FBI, CIA, IRS, the National Counterterrorism Center, local and state police agencies, and others without stripping names and personally identifying information, a practice that was approved by the FISC.

The tension, though, between the principles of civil rights transparency and effective public oversight and of effective national security practices is not a new one, and the tendencies of the information age will continue to evolve in these terms. It can be assured that the NSA will continue to be at the forefront of many of these controversies as the nation and the world decides where the appropriate legal boundary lies.

Cross-References

- [Ethical and Legal Issues](#)
- [Fourth Amendment](#)
- [Privacy](#)

Further Reading

- Bamford, J. (2014, August). Edward Snowden: The untold story. *WIRED*. <http://www.wired.com/2014/08/edward-snowden/>.
- Greenwald, G. (2014). *No place to hide: Edward Snowden, the NSA, and the U.S. surveillance state*. New York: Metropolitan Books.
- Macaskill, E., & Dance, G. (2013, November 1). NSA files: Decoded: What the revelations mean for you. *The Guardian*. <http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded>.
- National Security Administration. (2013). *60 years of defending our nation*. www.nsa.gov/about/crypto-logic_heritage/60th/book/NSA_60th_Anniversary.pdf.

Natural Disasters

► Natural Hazards

Natural Hazards

Guido Cervone¹, Yuzuru Tanaka² and Nigel Waters³

¹Geography, and Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, PA, USA

²Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan

³Department of Geography and Civil Engineering, University of Calgary, Calgary, AB, Canada

Synonyms

Disaster management; Natural disasters

Introduction

The origins of natural hazard events may be atmospheric/meteorological (droughts, heat waves, and storms such as cyclonic, ice, blizzards, hail, and tornados), hydrological (river and coastal

floods), geological (avalanches, coastal erosion, landslides, earthquakes, lahars, volcanic eruptions), and wildfires and extraterrestrial events (geomagnetic storms or impacts). These natural hazards, due to their location, severity, and frequency, may adversely affect humans, their infrastructure, and their activities. Climate change may exacerbate natural disasters due to weather by increasing the intensity and frequency of such disasters.

Research into natural hazards falls into four areas: mitigation, preparedness, response, and recovery, and this categorization will be followed here, even though the four areas often overlap. In all four areas, big data intensifies the challenges in carrying out these responses to natural hazards. Big data in these areas of research is impacted by the “seven Vs”: volume, variety, velocity, veracity, value, variability, and visualization (Akteer and Fosso Wamba 2017) to which may be added vinculation, viscosity, and vicinity. All these requirements of the big data that might be used in natural hazard research and operations would impact the demands on computational resources. Cloud computing is an active big data research area for natural hazards because it provides elastic computing to respond to varying computational loads that might occur in different geographical areas with differing probabilities of being affected during a disaster (Huang and Cervone 2016).

Major sources of research are the peer reviewed journals such as *International Journal of Disaster Risk Reduction*, *Journal of Disaster Research*, *Journal of Geography and Natural Disasters*, *Natural Hazards*, *Natural Hazards and Earth Systems Sciences*, *Natural Hazards Review*, and *Safety Science*. In each of the four areas of natural hazard, research reference will be made to recent case studies, although, as Akteer and Fosso Wamba (2017) have noted, these are much less common than review/conceptual or mathematical/analytical articles.

Natural Hazard Mitigation

Natural hazard mitigation measures are those undertaken by individuals or various levels of

government to reduce or eliminate the impacts of hazards or to remove the risk of damage and disaster. Various methodologies are used to assess the effectiveness of mitigation including return on investment (ROI). Although all natural hazards require mitigation, here this process will be illustrated by considering floods. Floods are among the most devastating of natural hazards including the two most deadly natural disasters of all time: the 1931 China floods that killed between 1 and four million people and the 1887 Yellow River Flood that killed between 900,000 and two million people. They also cause some of the most devastating environmental impacts (e.g., the Mozambique Flood of 2000 covered much of the country for about 3 weeks, an area of 1,400 sq. km). Floods are thus extensive and occur across the globe and with great and increasing frequency. All of this exacerbates the big data problems associated with their mitigation. Details of mitigation strategies for other natural hazards may be found in the FEMA Comprehensive Preparedness Guide (FEMA 2018).

Mitigation measures may be classed as structural or nonstructural. For example, in the case of a flood, traditional nonstructural approaches such as early detection and warning measures, zoning and building codes, emergency plans, and flood proofing and flood insurance may be supplemented by newer nonstructural approaches, to flood mitigation that include new computer architectures such as virtual databases and a decision support system to manage flood waters. Other methodological approaches are benefit-cost ratios (BCR) and cost-benefit analysis (CBA); a discussion of these can be found in Wisner et al. (2012).

Structural approaches to flood mitigation include, for example, the building of floodways to take flood waters away from residential areas. One of the most successful of these was the Red River Floodway in Manitoba, Canada. This floodway was originally built between 1962 and 1968 at a cost of Can\$63 million. Starting in 2005 a further Can\$627 million was spent to upgrade the capacity of the floodway from 90,000 to 140,000 cubic feet per second. It is estimated that it has prevented approximately Can\$12 billion worth of damage during major floods

to the Red River Basin and will protect up to a 1-in-700-year flood. Lee and Kim (2017) in a study of flood mitigation in Seoul, South Korea, describe an approach that combines structural and nonstructural flood prevention for decentralized reservoirs. They also review the extant literature for structural, nonstructural, and integrated approaches to flood mitigation.

Big data approaches to flood mitigation have been pioneered by the Dutch in collaboration with IBM using the Digital Delta software (Woodie 2013). This software allows the analysis of a huge variety of flood-related data. These data include water levels and water quality; sensors embedded in levees; radar data; and weather predictions and other historical flood-related information. To mitigate a flood, there is a need for evidence-based approaches, and this may be facilitated with the use of crowdsourcing and social media (Huang and Cervone 2016). This is a big data problem since all aspects of big data noted above add to the computational challenges faced by these researchers. Mitigation measures for all the other types of natural hazards may be found in Wisner et al. (2012).

Preparedness (Prevention and Protection)

FEMA (2018) provides a comprehensive guide to threat and hazard identification and risk assessment (THIRA). Their document describes five core capabilities: prevention, protection, mitigation, response, and recovery. Mitigation has been considered above, while response and recovery are considered separately below. Prevention and protection of threats and hazards fall into three areas: technological, human caused, and natural. Here the concern is only with the natural hazards listed in the introduction above. Preparedness is enhanced by developing and then consulting various sources such as emergency laws, policies, plans, and procedures, by checking existing THIRAs, and by planning emergency response scenarios with all levels of government, stakeholders, and first responders (fire, police, and emergency medical services).

Ideally these activities are coordinated with an emergency operations center (EOC). In addition, records and historical data from previous incidents should be reviewed and critical infrastructure interdependencies examined. Factors for selecting threats and hazards include the likelihood/probability of the incident and its significance in terms of impact. The complexity of these activities inevitably produces big data problems in all but the smallest communities. This is especially the case if a methodology needs to be developed for an all-hazards approach as opposed to the less demanding but more commonly encountered single hazard methodologies. Preparedness is most effective if the occurrence of a natural hazard can be predicted. Sala (2016) explains how hurricanes, minor seismic disturbances, and floods can be predicted using big data acquired from mobile phone accelerometers and crowdsourced from volunteers. These data may be collected using cloud computing and amalgamated with data from traditional seismic sources. Sala describes how researchers at the Quake-Catcher Network have gathered these data into the globally distributed Quake-Catcher Network that can be used as an early warning system for seismic disturbances, thus enhancing preparedness.

Response

Early detection systems blend into response systems. Koshimura (2017) has described a series of research initiatives under a Japan Science and Technology Agency, CREST, and big data applications program. These include a framework for the real-time simulation of a tsunami inundation that incorporates an estimation of building and other infrastructure damage; a study of the traffic distribution following the 2016 Kumamoto earthquake permitting the simulation of future traffic disruptions following similar natural disasters; the use of synthetic aperture radar (SAR) for damage detection following the 2016 Kumamoto, 2011 Great East Japan, and 2015 Nepal earthquakes; emergency vehicle and wide-area evacuation simulation models; a big data assimilation team to simulate the distribution of humans and cars assuming various scenarios following a natural disaster; and a

simulation of a big data warehouse for sharing the results of these simulations.

Hultquist and Cervone (2018) describe damage assessment of the urban environment during a natural disaster using various sources of volunteered geographic information (VGI). These sources include social media (Twitter, Facebook, Instagram to provide text, videos, and photos), mobile phones, collective mapping projects, and images from unmanned aerial vehicles (UAVs). The need for high levels of granularity in both space and time plus the integration of these interrelated (vinculation) VGI data with authoritative sources created big data demands for those analyzing and interpreting the information. The effectiveness of these data sources was proven with an analysis of the September 2013 floods in Colorado and health hazard monitoring following the 2011 Fukushima Daiichi nuclear disaster. Flood detection, warning, damage assessment, response, as well as disaster prevention and mitigation are all goals of the Dartmouth Flood Observatory (Sala 2016). United States Geological Survey (USGS) seismic data and National Aeronautics and Space Administration (NASA) Tropical Rainfall Measuring Mission (TRMM) rainfall data have been integrated with social sensors including YouTube, Instagram, and Twitter for landslide detection using the LITMUS system described in Sala (2016).

Tanaka et al. (2014) have addressed the problem of snow management in the city of Sapporo on the island of Hokkaido, Japan. Each year Sapporo with a population of almost two million receives approximately 6 m of snow and has a snow removal budget of almost \$180 million (US). Historic data from probe cars (private vehicles and taxis), buses, and snow plows are combined with real-time data from each of these sources. In addition, the system integrates (a vinculation big data problem) probe person data, traffic sensor data, meteorological sensor data, plus snow plowing and subway passenger records among other data sources. Visualization tools are integrated to minimize the impact of the snow hazard.

Recovery

An initial concern in disaster recovery is data restoration from an emergency operations center

or from affected businesses. Huang et al. (2017) review the literature on this and then describe how cloud computing can be used to rapidly restore large volumes of data to multiple operations centers. Business continuity refers to the restoration of IT or technology systems and the physical infrastructure of the environment damaged during the natural disaster.

FEMA (2016) has developed a National Disaster Recovery Framework (NDRF) that is designed to ensure not only the restoration of the community's physical infrastructure to pre-disaster conditions but also seeks to support the financial, emotional, and physical requirements of affected community members. The complexity of this task and the need for a rapid and integrated response to recovery ensure that this is a big data problem.

Conclusion

Natural hazards are the continuing source of disasters that impact communities around the world. Remediation of the threats that result from these hazards has been reviewed under the headings of mitigation, preparedness, response, and recovery. The complexity and interrelatedness of these tasks and the speed required for timely response ensure that they are "big data" problems. In the instances of atmospheric, meteorological, and hydrological events, these tasks are continuing to be exacerbated by climate change as extreme events become more frequent and of greater severity.

Further Reading

- Akter, S., & Fosso Wamba, S. (2017). Big data and disaster management: A systematic review and agenda for future research. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-017-2584-2>.
- FEMA. (2016). *National disaster recovery framework* (2nd ed.). Washington, DC: Federal Emergency Management Agency. 53 p.
- FEMA. (2018). Comprehensive preparedness guide (CPG) 201: Threat and hazard identification and risk assessment (THIRA) and Stakeholder preparedness review (SPR) guide. <https://www.fema.gov/media-library/assets/documents/165308>.

- Huang, Q., & Cervone, G. (2016). Usage of social media and cloud computing during natural hazards. In T. C. Vance, N. Merati, C. Yang, & M. Yuan (Eds.), *Cloud computing in ocean and atmospheric sciences* (pp. 297–324). Amsterdam: Academic Press.
- Huang, Q., Cervone, G., & Zhang, G. (2017). A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data. *Computers, Environment and Urban Systems*, 66:23–37. <https://doi.org/10.1016/j.compenvurbsys.2017.06.004>.
- Hultquist, C., & Cervone, G. (2018). Citizen monitoring during hazards: validation of Fukushima radiation measurements. *GeoJournal*, 83(2):189–206. <https://doi.org/10.1007/s10708-017-9767-x>.
- Koshimura, S. (2017). Fusion of real-time disaster simulation and big data assimilation – Recent progress. *Journal of Disaster Research*, 12(2), 226–232.
- Lee, E. H., & Kim, J. H. (2017). Design and operation of decentralized reservoirs in urban drainage systems. *Water*, 9, 246. <https://doi.org/10.3390/w9040246>.
- Sala, Simone 2016. Using big data to detect and predict natural hazards better and faster: Lessons learned with hurricanes, earthquakes and floods. <http://datapopalliance.org/using-big-data-to-detect-and-predict-natural-hazards-better-and-faster-lessons-learned-with-hurricanes-earthquakes-floods/>
- Tanaka, Y., Sjöbergh, J., Moiseets, P., Kuwahara, M., Imura, H., & Yoshida, T. (2014). Geospatial visual analytics of traffic and weather data for better winter road management. In G. Cervone, J. Lin, & N. Waters (Eds.), *Data mining for geoinformatics* (pp. 105–126). New York: Springer.
- Wisner, B., Gaillard, J. C., & Kelman, I. (Eds.). (2012). *Handbook of hazards and disaster risk reduction and management*. New York: Routledge.
- Woodie, Alex 2013. Dutch turn to big data for water management and flood control. https://www.datanami.com/2013/06/27/dutch_turn_to_big_data_for_water_management_flood_control/.

Natural Language Processing (NLP)

Erik W. Kuiler

George Mason University, Arlington, VA, USA

Natural Language Processing (NLP) – with Machine Learning (ML) and Deep Learning (DL) – constitutes an important subdomain of Artificial Intelligence (AI). NLP operates on very large unstructured data sets – text-based big data sets – by employing information technology

(IT) capabilities and linguistics to support computer-enabled Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLP provides not only the basis for text analyses of massive corpora, but also for such tools as virtual assistant AI technology (e.g., Siri and Alexa). Common applications of NLP include machine translating human languages into machine languages for analysis, manipulation, and management; supporting search engines; extracting and summarizing information from diverse sources (e.g., financial information from newspaper articles); supporting human-machine vocal interactions (e.g., Alexa and Siri); and filtering Spam. As a big data process, NLP can be framed and understood in basic terms referencing linguistic foundations, text analysis tasks, and text-based information extraction.

Linguistic Foundations

NLP begins with the application of linguistics – the scientific study of language – that comprises several disciplines. NLP uses linguistics to derive meaning from human speech and texts (referencing English examples):

Phonetics – the study of speech sounds and how they are made; for example, the sound *m* is articulated with the lips held closed; *b* starts with the lips held together, followed by a voiceless plosive.

Phonology – the study of distinguishable units of speech that distinguish one word from another – phonemes; for example, *b*, *p*, *h*: *bit*, *pit*, *hit*. NLP uses phonemes and their combination to identify comprehensible speech events based on predetermined lexica and usage conventions

Morphology – the study of how words are formed – morphemes, units of language that cannot be meaningfully be subdivided; for example, *out*, *go*, *-ing* collectively form *outgoing*; morphemes provide the basis for lexicon and ontology development. NLP uses morphemes to determine the construction and

potential roles of words. There are also lexical aspects: NLP examines how morphemes combine to make words and how minor differences can change the meaning of the word.

Syntax – the study of how sentences are formed and the rules that apply to their formulation; for example, a sentence may adopt a syntactic pattern of subject + verb+ direct object: *Louis hit the ball*. NLP uses predetermined syntactic rules and norms to determine the meaning of a sentence based on word order and its dependencies.

Semantics – the study of the meaning of language (not to be confused with semiotics – the study of symbols and their interpretations); for example, the subtle difference between *rubric* and *heading*. Based on the semantics of words and their syntax in a sentence, NLP attempts to determine what is the most likely meaning of a sentence and what makes the most sense in a specific context or discourse.

Pragmatics – the study of language and the circumstances in which it is used; for example, how people take turns in conversation, how texts are organized, how a word can take on a particular meaning based on tone or context; for example, *guilt* in a legal context differs from *guilt* in an ecclesiastical context. NLP uses semantics to determine how the contextual framework of a sentence helps determine the meaning of individual words.

NLP Text Analysis Tasks

Using NLP to perform text analysis usually takes the form several basic activities:

Sentence segmentation – demarcating separate sentences in a text.

Word tokenization – assigning tokens to each word in a sentence to make them machine-readable so that sentences can be processed one at a time.

Parts of speech assignment – designate a part of speech for each token (noun, pronoun, verb, adjective, adverb, preposition, conjunction,

interjection) in a sentence, facilitating syntactic conformance and semantic cohesion.

Text lemmatization – determining the basic form – *lemma* – of each word in a sentence; for example, *smok-* in *smoke*, *smoker*, or *smoking*.

Stop words identification – many languages have stop words, such as *and*, *the*, *a* in English; these are usually removed to facilitate NLP processing.

Dependency parsing – determining the relationships and dependencies of the words in a sentence; for example, noun phrases and verb phrases in a sentence.

Named entity recognition (NER) – examples of named entities that a typical system can identify are people's names, company names, physical and political geographic locations, product names, dates and times, currency amounts, and named events. NER is generally based on grammar rules and supervised models. However, there are NER platforms with built-in NER models.

Co-reference resolution – resolving the references of deictic pronouns such as *he*, *she*, *it*, *they*, *them*, etc.

Fact extraction – using a predefined knowledge base to extract facts (meaning) from a text.

NLP Text-based Information Extraction

NLP provides important capabilities to support diverse analytical efforts:

Sentiment analysis – NLP can be useful in analyses of people's opinions or feedback, such as those contained in customer surveys, reviews, and social media.

Aspect mining – aspect mining identifies different points of view (aspects) in a text. Aspect mining may be used in conjunction with sentiment analysis to extract information from a text.

Text summarization and knowledge extraction – NLP can be applied to extract information from, for example, newspaper articles or

research papers. NLP abstraction methods create summary by generating fresh text that conveys the crux of the original text; for example, an NLP extraction method can create a summary by extracting parts from the text.

Topic modelling – topic modeling focuses on identifying topics in a text and can be quite complex. An important advantage of topic modeling is that it is an unsupervised technique that does not require model training or a labeled training set. Algorithms that support topic modelling include the following:

Correlated Topic Model (CTM) – CTM is a topic model of a document collection that models the words of each document and correlates the different topics in the collection.

Latent Semantic Analysis (LSA) – an NLP technique for distributional semantics based on analyzing relationships between a set of documents and the terms they contain to produce a set of concepts related to the documents and terms.

Probabilistic Latent Semantic Analysis (PLSA) – A statistics-based technique for analyzing bi-model and co-occurrence data that can be applied to unstructured text data.

Latent Dirichlet Allocation (LDA) – The premise of LDA is that a text document in a corpus comprises topics and that each topic comprises several words. The input required by LDA is text documents and the number of topics that LDA algorithms are expected to generate.

Summary

NLP supports intellectual- and labour-intensive tasks, ranging from sentence segmentation and word tokenization to topic extraction and modeling. NLP is an important subdomain of AI, providing IT capabilities to analyze very large sets of unstructured data, such as text and speech data.

Further Reading

- Bender, E. M. (2013). *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. New York: Morgan & Claypool Publishers.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. Sebastopol: O'Reilly Publishing.
- Manning, C. D., & Schütze, H. (2002). *Foundations of natural language processing*. Cambridge, MA: MIT Press.
- Mitkov, R. (2003). *The Oxford book of computational linguistics*. Oxford: Oxford University Press.

Netflix

J. Jacob Jenkins
California State University Channel Islands,
Camarillo, CA, USA

Introduction

Netflix is a film and television provider headquartered in Los Gatos, California. Netflix was founded in 1997 as an online movie rental service, using Permit Reply Mail to deliver DVDs. In 2007, the company introduced streaming content, which allowed customers instant access to its online video library. Netflix has since continued its trend toward streaming services by developing a variety of original and award-winning programming. Due to its successful implementation of Big Data, Netflix has experienced exponential growth since its inception. It currently offers over 100,000 titles on DVD and is the world's largest on-demand streaming service with more than 80 million subscribers in over 190 countries worldwide.

Netflix and Big Data

Software executives Marc Randolph and Reed Hastings founded Netflix in 1997. Randolph was a previous cofounder of MicroWarehouse, a mail-order computer company; Hastings was a previous math teacher and founder of Pure Soft, a

software company he sold for \$700 million. The idea for Netflix was prompted by Hastings' experience of paying \$40 in overdue fees at a local Blockbuster. Using \$2.5 million dollars in start-up money from his sale of Pure Soft, Hastings envisioned a video provider whose content could be returned from the comfort of one's own home, void of due dates or late fees. Netflix's website was subsequently launched on August 29, 1997.

Netflix's original business model used a traditional pay-per-rental approach, charging 0.50 cents per film. Netflix introduced its monthly flat-fee subscription service in September 1999, which led to the termination of its pay-per-rental model by early 2000. Netflix has since built its global reputation on the flat-fee business model, as well as its lack of due dates, late fees, or shipping and handling charges. Netflix delivers DVDs directly to its subscribers using the United States Postal Service and a series of regional warehouses located throughout the United States. Based upon which subscription plan is chosen, users can keep between one and eight DVDs at a time, for as long as they desire. When subscribers return a disc to Netflix using one of its prepaid envelopes, the next DVD on their online rental queue is automatically mailed in its stead. DVD-by-mail subscribers can access and manage their online rental queue through Netflix's website in order to add and delete titles or rearrange their priority.

In 2007 Netflix introduced streaming content as part of its "Watch Instantly" initiative. When Netflix first introduced streaming video to its website, subscribers were allowed 1 h of access for every \$1 spent on their monthly subscription. This restriction was later removed due to emerging competition from Hulu, Apple TV, Amazon Prime, and other on-demand services. There are substantially less titles available through Netflix's streaming service than its disc library. Despite this limitation, Netflix has become the most widely supported streaming service in the world by partnering with Sony, Nintendo, and Microsoft to allow access through Blu-ray DVD players, as well as the Wii, Xbox, and PlayStation gaming consoles. In subsequent years, Netflix has increasingly turned attention toward its streaming

services. In 2008 the company added 2500 new “Watch Instantly” titles through a partnership with Starz Entertainment. In 2010 Netflix inked deals with Paramount Pictures, Metro-Goldwyn-Mayer, and Lions Gate Entertainment; in 2012 it inked a deal with DreamWorks Animation.

Netflix has also bolstered its online library by developing its own programming. In 2011 Netflix announced plans to acquire and produce original content for its streaming service. That same year it outbid HBO, AMC, and Showtime to acquire the production rights for *House of Cards*, a political drama based on the BBC miniseries of the same name. *House of Cards* was released on Netflix in its entirety in early 2013. Additional programming released during 2013 included *Lilyhammer*, *Hemlock Grove*, *Orange Is the New Black*, and the fourth season of *Arrested Development* – a series that originally aired on Fox between 2003 and 2006. Netflix later received the first Emmy Award nomination for an exclusively online television series. *House of Cards*, *Hemlock Grove*, and *Arrested Development* received a total of 14 nominations at the 2013 Primetime Emmy Awards; *House of Cards* received an additional four nominations at the 2014 Golden Globe Awards. In the end, *House of Cards* won three Emmy Awards for “Outstanding Casting for a Drama Series,” “Outstanding Directing for a Drama Series,” and “Outstanding Cinematography for a Single-Camera Series.” It won one Golden Globe for “Best Actress in a Television Series Drama.”

Through its combination of DVD rentals, streaming services, and original programming, Netflix has grown exponentially since 1997. In 2000, the company had approximately 300,000 subscribers. By 2005 that number grew to nearly 4 million users, and by 2010 it grew to 20 million. During this time, Netflix’s initial public offering (IPO) of \$15 per share soared to nearly \$500, with a reported annual revenue of more than \$6.78 billion in 2015. Today, Netflix is the largest source of Internet traffic in all of North America. Its subscribers stream more than 1 billion hours of media content each month, approximating one-third of total downstream web traffic. Such success has resulted in several competitors for online

streaming and DVD rentals. Wal-Mart began its own online rental service in 2002 before acquiring the Internet delivery network, Vudu, in 2010. Amazon Prime, Redbox Instant, Blockbuster @ Home, and even “adult video” services like WantedList and SugarDVD have also entered the video streaming market. Competition from Blockbuster sparked a price war in 2004, yet Netflix remains the industry leader in online movie rentals and streaming.

Netflix owes much of its success to the innovative use of Big Data. Because it is an Internet-based company, Netflix has access to an unprecedented amount of viewer behavior. Broadcast networks have traditionally relied on approximated ratings and focus group feedback to make decisions about their content and airtime. In contrast, Netflix can aggregate specified data about customers’ actual viewing habits in real time, allowing it to understand subscriber trends and tendencies at a much more sophisticated level. The type of information Netflix gathers is not limited to what viewers watch and the ratings they ascribe. Netflix also tracks the specific dates and times in which viewers watch particular programming, as well as their geographic locations, search histories, and scrolling patterns; when they use pause, rewind, or fast-forward; the types of streaming devices employed; and so on.

The information Netflix collects allows it to deliver unrivaled personalization to each individual customer. This customization not only results in better recommendations but also helps to inform what content the company should invest in. Once content has been acquired/developed, Netflix’s algorithms also help to optimize their marketing and to increase renewal rates on original programming. As an example, Netflix created ten distinct trailers to promote their original series *House of Cards*. Each trailer was designed for a different audience and seen by various customers based on those customers’ previous viewing behaviors. Meanwhile, the renewal rate for original programming on traditional broadcast television is approximately 35%; the current renewal rate for original programming on Netflix is nearly 70%.

As successful as Netflix's use of Big Data has been, the company strives to keep pace with changes in viewer habits, as well as changes in its own product. When the majority of subscribers used Netflix's DVD-by-mail service, for instance, those customers consciously added new titles to their queue. Streaming services demand a more instantaneous and intuitive process of generating future recommendations. In response to developments such as this, Netflix initiated the "Netflix Prize" in 2006: a \$1 million payout to the first person or group of persons to formulate a superior algorithm for predicting viewer preferences. Over the next 3 years, more than 40,000 teams from 183 countries were given access to over 100 million user ratings. BellKor's Pragmatic Chaos was able to improve upon Netflix existing algorithm by approximately 10% and was announced as the award winner in 2009.

Conclusion

In summation, Netflix is presently the world's largest "Internet television network." Key turning points in the company's development have included a flat-rate subscription service, streaming content, and original programming. Much of the company's success has also been due to its innovative implementation of Big Data. An unprecedented level of information about customers' viewing habits has allowed Netflix to make informed decisions about programming development, promotion, and delivery. As a result, Netflix currently streams more than 1 billion hours of content per month to over 80 million subscribers in 190 countries and counting.

Cross-References

- [Algorithm](#)
- [Apple](#)
- [Communications](#)
- [Data Streaming](#)
- [Entertainment](#)

- [Facebook](#)
- [Social Media](#)

Further Reading

- Keating, G. (2013). *Netflixed: The epic battle for America's eyeballs*. London: Portfolio Trade.
- McCord, P. (2014). How Netflix reinvented HR. *Harvard Business Review*. <http://static1.squarespace.com/static/5666931569492e8e1cdb5afa/t/56749ea457eb8de4eb2f2a8b/1450483364426/How+Netflix+Reinvented+HR.pdf>. Accessed 5 Jan 2016.
- McDonald, K., & Smith-Rowsey, D. (2016). *The Netflix effect: Technology and entertainment in the 21st century*. London: Bloomsbury Academic.
- Simon, P. Big data lessons from Netflix. *Wired*. Retrieved from <https://www.wired.com/insights/2014/03/big-data-lessons-netflix/>.
- Wingfield, N., & Stelter, B. (2011, October 24). How Netflix lost 800,000 members, and good will. *The New York Times*. <http://faculty.ses.wsu.edu/rayb/econ301/Articles/Netflix%20Lost%20800,000%20Members%20.pdf>. Accessed 5 Jan 2016.

Network Advertising Initiative

Siona Listokin

Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

The Network Advertising Initiative (NAI) is a self-regulatory association in the United States (US), representing third parties in online advertising, and is one of the oldest industry-led efforts focused on consumer data privacy and security. It was initially formed in 1999 following industry engagement with the Federal Trade Commission (FTC) and consisted of ten firms that covered 90% of the network advertising industry. Membership rosters and rules have fluctuated significantly since the NAI's formation, and it is useful to evaluate the organization's evolution rather than its performance in any single year. Today, the Initiative has about 100 participating firms. The NAI has received praise from the FTC as a leader in the self-regulatory community. However, many

critics point to a history of lax enforcement, ineffective consumer choice tools, and insufficient industry representation.

Initial Evolution of NAI

The FTC invited online advertisers to consider self-regulating the online profiling industry in 1999, in advance of a workshop on the subject. At the time, the FTC was concerned with the lack of transparency to consumers as to the involvement of ad networks while using the Web. The initial NAI agreement with the FTC was founded on the four principles of notice, choice, access, and security. Over time, data use/limitation and data reliability were added to the foundational principles. Notably, consumer choice over online tracking was based on an “opt-out” model for non-personally identifying information. In 2001, the NAI launched a Web form that allowed consumers to opt-out of participating firms’ data collection in a single site, but did not directly address the concern about lack of consumer knowledge.

While the NAI continued to grow its self-regulatory guidelines, within a few years many of the founding firms dropped out of the initiative, during a period that coincided with less FTC scrutiny and engagement in consumer privacy regulation. Only two companies, Avenue A and DoubleClick were full participating members in 2002; five other founding firms were listed as associate members that did not engage in online preference marketing and were not part of the opt-out web form. The NAI added third-party enforcement through TRUSTe at this time to improve credibility through their Watchdog Reports, though the company was also a participating member of the Initiative. TRUSTe’s public disclosure of complaints and resolutions became increasingly opaque, culminating in a total absence of public enforcement by the end of 2006. The lack of industry representation and credible enforcement led many privacy advocacy groups to declare the NAI a failed attempt at strong self-regulation.

Self-Regulatory Guidelines

In response to criticism over the NAI’s membership, enforcement and narrow definitions of consumer choice over advertising network data collection, along with a new FTC report on self-regulation in online behavioral advertising, the Initiative updated its self-regulatory guidelines at the end of 2008 and allowed for public comment. The new guidelines were notable for expanding the definition of online advertising as the industry evolved in the decade since its founding. In addition, NAI supported a new effort in consumer education, addressing the transparency concerns that persisted since its founding. A later update in 2013 added data transfer and retention restrictions to the core principles. In addition, NAI joined other major advertising organizations and trade associations in the Digital Advertising Alliance, which offered its own mechanism for opting out of interest-based advertisements via its AdChoices tool. NAI began regulating cross-app ads in 2016.

The NAI now includes about 100 companies as full members; associate memberships no longer exist. The Initiative emphasizes its industry coverage and notes that nearly all ads served on the Internet seen in the USA. involve the technology of NAI members. Compliance and enforcement are conducted by the NAI itself, and utilizes ongoing manual reviews of opt-out pages as well as an in-house scanner to check if opt-out choices are honored or if privacy policies have changed. In its 2018 Compliance Report, the NAI reported receiving almost 2000 consumer and industry complaints, the vast majority of which were either outside of NAI’s mission or related to technical glitches in the opt-out tool. NAI investigating one potential noncompliance in 2018.

Assessment of NAI

There have been a number of outside assessments of the NAI following its 2008 update. It is worth noting that some of these evaluations are conducted by privacy advocacy groups that are

skeptical of self-regulation in general and supportive of comprehensive consumer privacy legislation in the USA. That said, the NAI is frequently criticized for inadequate technical innovation in its consumer choice tools and a lack of credible enforcement.

Despite general approval over the 2008 and 2013 updates, critiques of the NAI note that the main opt-out function has remained largely static, utilizing web-based cookies despite changing technologies and consumer behavior. In addition, the Initiative defines online behavioral advertising as that done by a third party, and its principles therefore do not apply to tracking and targeting by websites in general. A 2011 study found more than a third of NAI members did not remove their tracking cookies after the opt-out choice was selected in the Initiative's web form. Other works have found that only about 10% of those studied could discern the functionality of the NAI opt-out tool, and that there was infrequent compliance with membership requirements in both privacy policies and opt-out mechanisms. These studies also note the variability in privacy policy and opt-out options, with many membership firms going above and beyond the NAI code.

Further Reading

- Dixon, P. (2007). The network advertising initiative: Failing at consumer protection and at self-regulation. *World Privacy Forum*, Fall 2007.
- King, N. J., & Jessen, P. W. (2010). Profiling the mobile customer—Is industry self-regulation adequate to protect consumer privacy when behavioural advertisers target mobile phones? – Part II. *Computer Law & Security Review*, 26(6), 595–612.
- Komanduri, S., Shay, R., Norcie, G., Ur, B., & Cranor, L. F. (2011). AdChoices? Compliance with online behavioral advertising notice and choice requirements. *Carnegie Mellon University CyLab*, March 30, 2011.
- Mayer, J. (2011). Tracking the trackers: Early results. *Stanford Law School Center for Internet and Society*, July 12, 2011.

Network Analysis

► [Link/Graph Mining](#)

Network Analytics

Jürgen Pfeffer

Bavarian School of Public Policy, Technical University of Munich, Munich, Germany

Synonyms

[Network science](#); [Social network analysis](#)

Much of big data comes with relational information. People are friends with or follow each other on social media platforms, send each other emails, or call each other. Researchers around the world copublish their work, and large-scale technology networks like power grids and the Internet are the basis for worldwide connectivity. Big data networks are ubiquitous and are more and more available for researchers and companies to extract knowledge about our society or to leverage new business models based on data analytics. These networks consist of millions of interconnected entities and form complex socio-technical systems that are the fundamental structures governing our world, yet defy easy understanding. Instead, we must turn to network analytics to understand the structure and dynamics of these large-scale networked systems and to identify important or critical elements or to reveal groups. However, in the context of big data, network analytics is also faced with certain challenges.

Network Analytical Methods

Networks are defined as a set of nodes and a set of edges connecting the nodes. The major questions for network analytics, independent from network size, are “Who is important?” and “Where are the groups?” Stanley Wasserman and Katherine Faust have authored a seminal work on network analytical methods. Even though this work was published in the mid-1990s, it can still be seen as the standard book on methods for network analytics, and it also provides the foundation for many contemporary methods and metrics. With respect

to identifying the most important nodes in a given network, a diverse array of centrality metrics have been developed in the last decades. Marina Henning and her coauthors classified centrality metrics into four groups. “Activity” metrics purely count the number or summarize the volume of connections. For “radial” metrics, a node is important if it is close to other nodes, and “medial” metrics account for being in the middle of flows in networks or for bridging different areas of the network. “Feedback” metrics are based on the idea that centrality can result from the fact that a node is connected (directly or even indirectly) to other central nodes. For the first three groups, Linton C. Freeman has defined “degree centrality,” “closeness centrality,” and “betweenness centrality” as the most intuitive metrics. These metrics are used in almost every network analytical research project nowadays. The fourth metric category comprises mathematically advanced methods based on eigenvector computation. Phillip Bonacich presented eigenvector centrality which led to important developments of metrics for web analytics like Google’s PageRank algorithm or the HITS algorithms by John Kleinberg, which is incorporated into several search engines to rank search results based on the website’s structural importance on the Internet.

The second big pile of research questions related to networks is about identifying groups. Groups can refer to a broad array of definitions, e.g., nodes sharing of certain socioeconomic attributes, membership affiliations, or geographic proximity. When analyzing networks, we are often interested in structurally identifiable groups, i.e., sets of nodes of a network that are denser connected among them and sparser connected to all other nodes. The most obvious group of nodes in a network would be a clique – a set of nodes where each node is connected to all other nodes. Other definitions of groups are more relaxed. *K*-cores are a set of nodes for which every node is connected to at least *k* other nodes in the set. It turns out that *k*-cores are more realistic for real-world data than cliques and much faster to calculate. For any form of group identification in networks, we are often interested in evaluating the “goodness” of the identified groups. The most

common approach to assess the quality of grouping algorithms is to calculate the modularity index developed by Michelle Girvan and Mark Newman.

Algorithmic Challenges

The most widely used algorithms in network analytics were developed in the context of small groups of (less than 100) humans. When we study big networks with millions of nodes, several major challenges emerge. To begin with, most network algorithms run in $\Theta(n^2)$ time or slower. This means that if we double the number of nodes, the calculation time is quadrupled. For instance, let us assume we have a network with 1,000 nodes and a second network with one million nodes (thousandfold). If a certain centrality calculation with quadratic algorithmic complexity takes 1 min on the first network, the same calculation would take 1 million minutes (approximately 2 years) on the second network (millionfold). This property of many network metrics makes it nearly impossible to apply them to big data networks within reasonable time. Consequently, optimization and approximation algorithms of traditional metrics are developed and used to speed up analysis for big data networks.

A straight forward approach for algorithmic optimization of network algorithms for big data is parallelization. The abovementioned algorithms closeness and betweenness centralities are based on all-pairs shortest path calculation. In other words, the algorithm starts at a node, follows its links, and visits all other nodes in concentric circles. The calculation for one node is independent from the calculation for all other nodes; thus, different processors or different computers can jointly calculate a metric with very little coordination overhead.

Approximation algorithms try to estimate a centrality metric based on a small part of the actual calculations. The calculations of the all-pairs shortest path calculation can be restricted in two ways. First, we can limit the centrality calculation to the *k*-step neighborhood of nodes, i.e., instead of visiting all other nodes in concentric

circles, we stop at a distance k . Second, instead of all nodes, we just select a small proportion of nodes as starting points for the shortest path calculations. Both approaches can speed up calculation time tremendously as just a small proportion of the calculations are needed to create these results. Surprisingly, these approximated results have very high accuracy. This is because real-world networks are far from random and have specific characteristics. For instance, networks created from social interactions among people often have core-periphery structure and are highly clustered. These characteristics facilitate the accuracy of centrality approximation calculations. In the context of optimizing and approximating traditional network metrics, a major future challenge will be to estimate time/fidelity trade-offs (e.g., develop confidence intervals for network metrics) and to build systems that incorporate the constraints of user and infrastructure into the calculations. This is especially crucial as certain network metrics are very sensitive and small data change can lead to big change of results.

New algorithms are especially developed for very large networks. These algorithms have sub-quadratic complexity so that they are applicable for very large networks. Vladimir Batagelj and Andrej Mrvar have developed a broad array of new metrics and a network analytical tool called “Pajek” to analyze networks with tens of millions of nodes.

However, some networks are too big to fit into the memory of a single computer. Imagine a network with 1 billion nodes and 100 billion edges – social media networks have already reached this size. Such a network would require a computer with about 3,000 gigabyte RAM to hold the pure network structure with no additional information. Even though supercomputer installations already exist that can cope with these requirements, they are rare and expensive. Instead, researchers make use of computer clusters and analytical software optimized for distributed systems, like Hadoop.

Streaming Data

Most modern big data networks come from streaming data of interactions. Messages are sent

among nodes, people call each other, and data flows are measured among servers. The observed data consist of dyadic interaction. As the nodes of the dyads overlap over time, we can extract networks. Even though networks extracted from streaming data are inherently dynamic, the actual analysis of these networks is often done with static metrics, e.g., by comparing the networks created from daily aggregation of data. The most interesting research questions with respect to streaming data are related to change detection. Centrality metrics for every node or network level indices that describe the structure of the network can be calculated for every time interval. Looking at these values as time series can help to identify structural change in the dynamically changing networks over time.

Visualizing Big Data Networks

Visualizing networks can be a very efficient analytical approach as human perception is capable of identifying complex structures and patterns. To facilitate visual analytics, algorithms are needed that present network data in an interpretable way. One of the major challenges for network visualization algorithms is to calculate the positions of the nodes of the network in a way that it reveals the structure of the network, i.e., show communities and put important nodes in the center of the figure. The algorithmic challenges for visualizing big networks are very similar to the ones discussed above. Most commonly used layout algorithms scale very poorly. Ulrich Brandes and Christian Pich developed a layout algorithm based on eigenvector analysis that can be used to visualize networks with millions of nodes. The method that they applied is similar to the before-mentioned approximation approaches. As real-world networks normally have a certain topology that is far from random, calculating just a part of the actual layout algorithm can be a good enough approximation to reveal interesting aspects of a network.

Networks are often enriched with additional information about the nodes or the edges. We often know the gender or the location of people. Nodes might represent different types of

infrastructure elements. We can incorporate this information by mapping data to visual elements of our network visualization. Nodes can be visualized with different shapes (circles, boxes, etc.) and can be colored with different colors resulting in multivariate network drawings. Adding contextual information to compelling network visualizations can make the difference between pretty pictures and valuable pieces of information visualization.

Methodological Challenges

Besides algorithmic issues, we also face serious conceptual challenges when analyzing big data networks. Many “traditional” network analytical metrics were developed for groups of tens of people. Applying the same metrics to very big networks raises questions whether the algorithmic assumptions or the interpretations of results are still valid. For instance, the abovementioned metrics closeness and betweenness centralities just incorporate the shortest paths between every pair of nodes ignoring possible flow of information on non-shortest paths. Even more, these metrics do not take path length into account. In other words, if a node is on the shortest path of length, two or eight is treaded identically. Most likely this does not reflect real-world assumptions of information flow. All these issues can be addressed by applying different metrics that incorporate all possible paths or a random selection of paths with length k . In general, when accomplishing network analytics, we need to ask which of the existing network algorithms are suitable under which assumptions to be used for very large networks? Moreover, what research questions are appropriate for very large networks? Does being a central actor in a group of high school kids has the same interpretation as being a central user of an online social network with millions of users?

Conclusions

Networks are everywhere in big data. Analyzing these networks can be challenging. Due of the very nature of network data and algorithms, many

traditional approaches of handling and analyzing these networks are not scalable. Nonetheless, it is worthwhile coping with these challenges. Researchers from different academic areas have been optimizing existing and developing new metrics and methodologies as network analytics can provide unique insights into big data.

Cross-References

- ▶ [Algorithmic Complexity](#)
- ▶ [Complex Networks](#)
- ▶ [Data Streaming](#)
- ▶ [Data Visualization](#)

Further Reading

- Batagelj, V., Mrvar, A., & de Nooy, W. (2011). *Exploratory social network analysis with Pajek*. (Expanded edition.). New York: Cambridge University Press.
- Brandes, U., & Pich, C. (2007). Eigensolver Methods for progressive multidimensional scaling of large data. Proceedings of the 14th International Symposium on Graph Drawing (GD’06), 42–53.
- Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3), 215–239.
- Hennig, M., Brandes, U., Pfeffer, J., & Mergel, I. (2012). *Studying social networks. A guide to empirical research*. Frankfurt: Campus Verlag.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Network Data

Meng-Hao Li

George Mason University, Fairfax, VA, USA

Network (graph) data consist of composition and structural variables. Composition variables measure actor attributes, where actor attributes could be gender or age for people, private or public for organizations, country names, or locations. Structural variables measure connections between pairs of actors, where connections could be friendships between people, collaboration between organizations, trade between nations, or transmission line

between two stations (Wasserman and Faust 1994, p. 29; Newman 2010, p. 110). In the mathematical literature, network data are called as *graph* $G = (V, E)$, where V is the set of vertices (actors), and E is the set of edges (connections). Table 1 shows some examples of composition and structural variables in different properties of networks.

Modes of Networks

The *mode* of a network is used to express the number of sets of vertices on which structural variables present. There is no limitation on the number of the *mode* that a network can construct, but most networks are defined as either one-mode networks or two-mode networks. One-mode networks have one set of vertices that are similar to each other. For example, in a friendship network, the set of vertices is people connected by friendships. In Fig. 1, the friendship network has six vertices (1, 2, 3, 4, 5, 6) and seven edges (1, 2), (1, 5), (2, 4), (2, 5), (3, 4), (3, 5), and (5, 6). The representation of vertices and edges is also called an *edge list*. Edge lists are commonly used to store network data on computers and are often efficient for computing a large network.

Two-mode (affiliation; bipartite) networks consist of two sets of vertices. For example, a group of doctors work for several hospitals. Some doctors work for a same hospital but some doctors work for different hospitals. In this case, one set of vertices is the doctors and another set of vertices is the hospitals. In Fig. 2, the doctor-hospital network consists two sets of vertices, five doctors (a, b, c, d, e), and three hospitals (A, B, C). The edge represents that a doctor is affiliated to a hospital. For example, doctor b is

affiliated to hospital A, and doctor c is affiliated to hospitals A, B, and C.

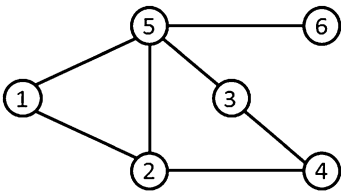
The Adjacency Matrix

The adjacency matrix is a most common form to represent a network mathematically. For example, the adjacency matrix of the friendship network in Fig. 1 can be displayed as elements A_{ij} in Table 2, where 1 represents that there is an edge between vertices i and j , and 0 represents that there is no edge between vertices i and j . This is a symmetric matrix with no self-edges, implying that the elements in the upper right and lower left triangles are identical, and all diagonal matrix elements are zero.

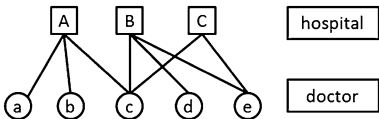
In Fig. 1, the friendship network that has single edge and no self-edge is also called *simple graph*. In some situations, a network may have multiple edges between two vertices (*multiedge* is also called *multiplexity* in Sociology). The network is called *multigraph*. Figure 3 is a representation of *multigraph*. Suppose that a researcher is interested in understanding a group’s friendship network, advice network, and gossip network. The researcher conducts a network survey to investigate how those people behave in those three networks. The survey data can be constructed as a *multigraph* network in Fig. 3. In Fig. 3, a solid edge represents a friendship, a dotted edge represents an advice relation, and a dash-dot edge represents a gossip relation. For example, there are friendship, advice, and gossip relations between vertices 1 and 5. The vertices 5 and 3 are connected by friendship and advice relations. The vertices 2 and 4 are linked by friendship and gossip relations. The *multigraph* network can also be converted to an adjacency matrix A_{ij} in Table 3.

Network Data, Table 1 Examples of networks

Network	Composition variable		Structural variable
	Vertex	Attribute	Edge
Friendship Network	Person	Age, gender, weight, or income	Friendship
Collaboration Network	Organization	Public, private, or nonprofit	Collaboration
Citation Network	Article	Biology, Engineering or Sociology	Citation
World Wide Web	Web page	Government, education or commerce	Hyperlink
Trade Network	Nation	Developed or developing	Trade



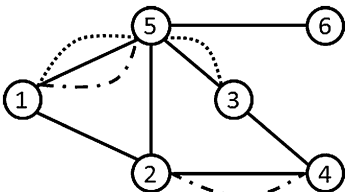
Network Data, Fig. 1 One-mode friendship network



Network Data, Fig. 2 Two-mode doctor-hospital network

Network Data, Table 2 Adjacency matrix

	1	2	3	4	5	6
1	0	1	0	0	1	0
2	1	0	0	1	1	0
3	0	0	0	1	1	0
4	0	1	1	0	0	0
5	1	1	1	0	0	1
6	0	0	0	0	1	0



Network Data, Fig. 3 An example of the multigraph network

Network Data, Table 3 The adjacency matrix of the multigraph network

	1	2	3	4	5	6
1	0	1	0	0	3	0
2	1	0	0	2	1	0
3	0	0	0	1	2	0
4	0	2	1	0	0	0
5	3	1	2	0	0	1
6	0	0	0	0	1	0

The values between i and j represent that the number of edges is present between i and j .

The Incidence Matrix

The incidence matrix is used to represent a two-mode (affiliation; bipartite) network. In Fig. 2, the two-mode doctor-hospital network can be constructed as an incidence matrix as B_{ij} in Table 4, where 1 represents doctor j belongs to hospital i , and 0 represents doctor j does not belong to hospital i . Although an incidence matrix can completely represent a two-mode network, it is often convenient for computing by projecting a two-mode network to a one-mode network. Tables 5 and 6 are different ways to exhibit the doctor-hospital network in one-mode networks. Table 5 is a hospital adjacency matrix, where 1 represents that there is at least one shared doctor between hospitals. Table 6 is a doctor adjacency matrix, where 1 represents that there is at least one shared hospital between doctors and 0 represents there is no shared hospital between doctors.

Network Data, Table 4 Incidence matrix

	a	b	c	d	e
A	1	1	1	0	0
B	0	0	1	0	1
C	0	0	1	1	1

Network Data, Table 5 Hospital adjacency matrix

	A	B	C
A	0	1	1
B	1	0	1
C	1	1	0

Network Data, Table 6 Doctor adjacency matrix

	a	b	c	d	e
a	0	1	1	0	0
b	1	0	1	0	0
c	1	1	0	1	1
d	0	0	1	0	1
e	0	0	1	1	0

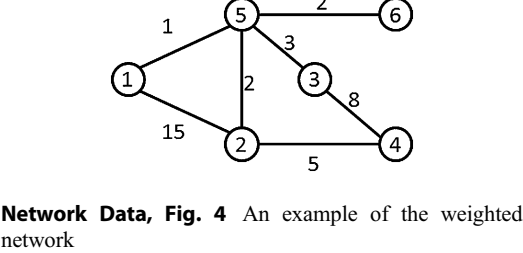
Weighted Networks

The aforementioned examples assume that edges are weighted equally, but it may not be realistic in most network structures. The weighted (valued) networks relax the edge value and allow a researcher to assign values for each edge in a network. For example, Fig. 1 has equal weights on each edge in the friendship network. The friendship can be weighted by the time that two people have known each other. Figure 4 shows the weighted friendship network of Fig. 1. The edge value 15 between vertices 1 and 2 represents that the two people have known each other for 15 years. Likewise, the edge 1 between vertices 1 and 5 represents that the two people have known each other for 1 year. Table 7 converted the weighted network into an adjacency matrix. This is a symmetric matrix with no self-edges. The upper right and lower left triangles have the same elements, and diagonal matrix elements are all zero.

Signed Networks

The signed networks are used to represent a network with “positive” and “negative” edges. A

negative edge does not refer to an absence of an edge. For example, Fig. 5 shows a network with positive friendship edges and negative animosity edges. A negative edge here represents that two enemies are connected by an animosity edge. A positive edge represents that two friends are connected by a friendship edge. If edges are absent between two people, it simply indicates that the two people do not interact with each other. The signed networks are commonly stored as two distinct networks, one with positive edges and the other one with negative edges. The adjacency matrix in Table 8 shows positive edges of the friendship network in Fig. 5, where 1 represents that there is a positive edge between two people and 0 represents that there is no edge between two people. The adjacency matrix in Table 9 is constructed by negative edges, where 1 represents that there is a negative edge between



Network Data, Table 7 The adjacency matrix of the weighted network

hh	1	2	3	4	5	6
1	0	15	0	0	1	0
2	15	0	0	5	2	0
3	0	0	0	8	3	0
4	0	5	8	0	0	0
5	1	2	3	0	0	2
6	0	0	0	0	2	0

Network Data, Fig. 5 An example of the signed network

Network Data, Table 8 Positive signed adjacency matrix

	1	2	3	4	5	6
1	0	1	0	0	1	0
2	1	0	0	0	1	0
3	0	0	0	1	0	0
4	0	0	1	0	0	0
5	1	1	0	0	0	0
6	0	0	0	0	0	0

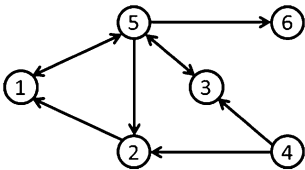
Network Data, Table 9 Negative signed adjacency matrix

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	0	1	0	0
3	0	0	0	0	1	0
4	0	1	0	0	0	0
5	0	0	1	0	0	1
6	0	0	0	0	1	0

two people and 0 represents that there is no edge between two people.

Directed Networks

A directed network is a network with directed edges, where an arrow points a direction from one vertex to another vertex. Suppose that Fig. 6 is a directed network consisting of directed friendships, indicating that some people recognize other people as their friends. For example, there is one directional edge from person 2 to person 1, indicating that person 2 recognizes that person 1 is her friend. But it does not mean that person 1 also recognizes that person 2 is her friend, the friendship between person 1 and person 2 is one direction and asymmetric. In the friendship between person 1 and person 5, the bi-directional arrow edge represents that there is a mutual recognition of the friendship between person 1 and person 5. As an example, the friendship network can be presented as an adjacency matrix A_{ij} in Fig. 6, where 1 represents that there is an edge from j to i , and 0 represents that there is no edge between j and i . It must be remembered, by convention and mathematical calculation, the direction goes from rows (j) to columns (i) (Table 10).



Network Data, Fig. 6 An example of the directed network

Network Data, Table 10 The adjacency matrix of the directed network

	1	2	3	4	5	6
1	0	1	0	0	1	0
2	0	0	0	1	1	0
3	0	0	0	1	1	0
4	0	0	0	0	0	0
5	1	0	1	0	0	0
6	0	0	0	0	1	0

Quality of Network Data

The quality of network data is principally determined by the methods of data collection and data cleansing. However, there is no universal method that can ensure a high quality of data. The only way that a researcher can pursue is to minimize threat of data errors and to seek optimal methods to approach research questions. Some types of errors have significant impacts on the quality of data and data analysis. Those errors are summarized below (Borgatti et al. 2013, pp. 37–40).

1. Omission errors: this type of errors describes missing edges or vertices in the data collection process. In Fig. 1, for example, vertex 5 has four connections with other vertices and seems to occupy an important position in the network. If information of vertex 5 was not collected, it would cause a significant bias in data analysis.
2. Commission errors: this type of errors describes that some edges or vertices should not be included in the network. In other words, network boundaries need to be set precisely to exclude unnecessary vertices and edges.
3. Edge/node attribution errors: this type of errors describes that attributes of edges or nodes are incorrectly assigned. For example, a private organization is labeled as a public organization in Table 1.
4. Retrospective errors: this type of errors describes that informant/respondent competence is not capable of recalling people and activities that they have involved in. This is an important issue in network survey studies. Since network survey questions are not intuitive as conventional survey questions, some respondents are burdened with identifying people and recognizing activities with those people (Marsden 2005, pp. 21–23).
5. Data management errors: this type of errors describes coding errors, inappropriate survey instruments or software issues.
6. Data aggregation errors: this type of errors describes information lost during the data aggregation process. It sometime occurs with omission errors. As an example, when a researcher needs to aggregate different data

sets into a master file, some vertices that do not match others may need to be excluded from the master file.

7. Errors in secondary data sources: this type of errors describes that data sources have inherent errors. For example, Facebook only allows researchers to access individual accounts that are open to public. Private accounts thus would be excluded from the data collection.
8. Formatting errors: this type of errors describes that data formatting process may cause errors. For example, a researcher collects data from different sources with various formats. When there is a need for the researcher to integrate various data formats for analysis, the formatting errors are likely to occur.

Large-Scale Network Data

Storing, querying, and analyzing network data have become extremely challenging when the scale of network data is large. In a conventional dataset, 1,000 data points generally represent 1,000 data points. In a network dataset, 1,000 homogenous vertices are likely to be connected by 499,500 undirected edges ($N \times (N-1)/2$). When network data come with heterogenous vertices, directed/weighted/signed edges, time points, and locations, the possible combinations of network structures will exponentially grow up (e.g., mobile cellular networks). Several graph database systems, graph processing systems, and graph dataflow systems have been developed to manage network data. Those systems typically require the allocation of distributed clusters and in-memory graph processing and are anticipated to offer flexible and efficient ways of querying and analyzing large-scale network data (Junghanns et al. 2017).

Cross-References

- [Big Variety Data](#)
- [Collaborative Filtering](#)
- [Data Aggregation](#)

- [Data Brokers and Data Services](#)
- [Data Cleansing](#)
- [Data Integration](#)
- [Data Quality Management](#)
- [Database Management Systems \(DBMS\)](#)

Further Reading

- Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks*. London, England: SAGE.
- Junghanns, M., Petermann, A., Neumann, M., & Rahm, E. (2017). Management and analysis of big graph data: Current systems and open challenges. In A. Y. Zomaya & S. Sakr (Eds.), *Handbook of big data technologies* (pp. 457–505). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-49340-4_14.
- Marsden, P. V. (2005). Recent Developments in Network Measurement. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis*. Cambridge: Cambridge University Press.
- Newman, M. (2010). *Networks: An introduction*. Oxford, England: Oxford University Press.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, England: Cambridge University Press.

Network Science

- [Link/Graph Mining](#)
 - [Network Analytics](#)
-

Neural Networks

Alberto Luis García
Departamento de Ciencias de la Comunicación
Aplicada, Facultad de Ciencias de la Información,
Universidad Complutense de Madrid, Madrid,
Spain

Neural networks are analytic techniques modeled inspired in the processes of learning by an animal's central nervous systems which is capable of

predicting new observations (on specific variables) from other observations (on the same or other variables).

Neural networks have seen an explosion of interest over the last few years and are aimed to apply across finance, medicine, research, classification, data processing, robotics, engineering, geology and physics, to get faster network processing, more efficiency, or fewer errors.

The two main characteristics of neural networks are *nonlinear*, i.e., *there is a possibility to introduce a large number of variables; also, neural networks are easy to use because works with training algorithms* to automatically learn the structure of the data. Neural networks are also intuitive, based on the similarity with the biological neural systems.

Neural networks have grown out of research in artificial intelligence, with which its structure based research on the development of knowledge of brain functioning. The main branch of artificial intelligence research in the 1960s–1980s produced expert systems.

The brain is composed of 10,000,000,000 of *neurons*, massively interconnected between them. Each neuron is a specialized cell, composed of dendrites (input structure) and axon (output structure and connect with dendrites of another neuron via a synapse).

Work through neural networks is structured around two main characteristics: the size of the structure and the number of layers needed to meet all the variables specified in the model. In all models, the main form of work is through trial and error testing.

The new network is then subjected to the process of training and learning, where it is applied an iterative process of inputs variables adjusted to the weights of the network in order to optimally predict the sample data. The Network developed in this process is a pattern that can make predictions through real input data, but that can be modified through the different layers adjusting the results in a specific data.

One of the major advances in working with neural networks is to prevent initial working

hypotheses, which can lead to erroneous research lines go. It is the very model which will suggest trends and at the same time, learns that run through the intervening variables to continuously adapt to circumstances. An important disadvantage, however, is that the final solution depends on the process of the initial training and learning and the initial conditions of the network.

The main applications for neural networks are data mining and exploratory data analysis, but there exist another one. Neural networks can be applied in all the situations that consist in a relationship between the predictor variables (inputs) and predicted variables (outputs).

For example:

- **Detection of medical phenomena.** These models are used as a way to prevent major pest or disease control in large populations. Also, to monitor and prevent the processes of disease development and, thus, to make tighter nursing diagnosis.
- **Stock market prediction.** Nowadays, it is very important to know the fluctuations of DOW, NASDAQ, or FTSE index, and try to predict the tomorrow's stock prices. In some circumstances, there are partially deterministic phenomenons (as factors such as past performance of other stocks and various economic indicators), that can be used to learn the model.
- **Credit assignment.** There are always the same variables to analyze the risk of a credit (the applicant's age, education, occupation, . . .). It is possible to work with all of this variables as inputs in the model and with the previous credit history, analyze and predict the risk in a specifically way for each client.
- **Monitoring the machinery.** Neural networks can be to be used in the preventive maintenance of machines, trained to distinguish between good and bad performance of a machine.
- **Engine management.** Neural networks have been used to analyze the input of sensors from an engine, i.e., in a Formula 1 race. All sensors are monitored machine and create a historical work that allows you to train and teach the model to make predictions naturally very

reliable operation. This way we can avoid additional costs in repairs and maintenance of engines and machines.

- **Image processing.** With proper training, you are able to read a car license plate or recognize a person’s face.

The main question is how we apply to solve a problem with a Neural Network. The first thing is to apply the model to a specific problem that can be solved through historical data, repetitive situations, etc. For example, it is impossible to predict the lottery (in a normal way). Another important requirement is that there are relationship variables between inputs and outputs data; the relationship can be strange, but it must exist. And more, it can be possible to begin with the model in training and a learning process that can be supervised or unsupervised. In the first, the training data contains examples of input data (as historical data of events, historical fluctuations of stock process, etc.) that are controlled by the researcher and examples of output data. The results are adjusted to the model and it can be known the final result to check the success of the model.

If the model is “ready” to work, the first decision is to choose which variables to use and how many cases to gather. The choice of variables is guided by intuition and the process of work with this chosen variables can be determined to choose another ones. But the first part of the process is the choice of the main influential variables in the process. This data can be numeric that must be scaled into an appropriate range for the network and another kind of statistic values.

Classification of neural networks according to network topology	
Monolayer neural network	It is the simplest neural network and is formed by an input neuron layer and an output neuron layer
Multilayer neural network	A neural network formed by several layers, in which there are hidden intermediate layers between the input layer and the output layer

(continued)

Convolutional neural network (CNN)	It is a multilayer network in which each part of the network is specialized to perform a task, thus reducing the number of hidden layers and allowing faster training
Recurrent neural network (RNN)	These are networks without a layered structure but allow for arbitrary connections between neurons
Radial-based network (RBF)	Calculates the output of the function according to the distance to a point called the center, avoiding local information minima where information back-propagation may be blocked
Classification of neural networks according to the learning method	
Supervised learning	Supervised learning is learning based on the supervision of a controller who evaluates and modifies the response according to its correctness or falsity
Error correction learning	Adjusts values according to the difference between expected and obtained values
Stochastic learning	Uses random changes that modify the weights of variables, keeping those that improve the results
Unsupervised or self-supervised learning	The algorithm itself and the internal rules of the neural networks create consistent output patterns. The interpretation of the data depends on the learning algorithm used
Hebrew Learning	Measures the familiarity and characteristics of the input data
Competitive and comparative learning	It consists of adding data and leaving out those that are more similar to the input pattern, giving more weight to those data that meet this premise
Learning by reinforcement	It is similar to supervised learning but it is only indicated if the data is acceptable or not

The current predisposition and enthusiasm towards artificial intelligence is largely due to advances in deep learning, which is based on techniques that allow the implementation of automatic learning of the algorithms that build and determine artificial neural networks. Deep learning is based on interaction based on the functioning of the human brain, in which several layers of interconnected simulated neurons learn to understand more complex processes. Deep learning networks have more than ten layers with millions of neurons each
Deep learning is made possible by the Big Data to train and teach systems, storage capacity and system performance, both in terms of storage and developers of cores, CPUs, and graphics cards

Further Reading

Boonkiatpong, K., & Sinthupinyo, S. Applying multiple neural networks on large scale data. In *2011 International Conference on Information and Electronics Engineering, IPCSIT*, Vol. 6. Singapore: Press.

NoSQL (Not Structured Query Language)

Rajeev Agrawal¹ and Ashiq Imran²

¹Information Technology Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, USA

²Department of Computer Science & Engineering, University of Texas at Arlington, Arlington, TX, USA

Synonyms

Big Data; Big Data analytics; Column-based database; Document-oriented database; Key-value-based database

Introduction

Rapidly growing humongous amount of data must be stored into a database. NoSQL is increasingly used in order to store Big Data. NoSQL systems are also called “not only SQL” or “not relational SQL” to emphasize that they may also support SQL query but more than that. Moreover, a NoSQL data store is able to accept all types of data – structured, semi-structured, and unstructured – much more easily than a relational database. For applications that have a mixture of datatypes, a NoSQL database is a good option. Performance factors come into play with an RDBMS’ data model, especially where “wide rows” are involved and update actions are many. However, a NoSQL data model such as Google’s Bigtable easily

handles both situations and delivers very fast performance for both read and write operations. NoSQL database addresses the following opportunities (Cattell 2011):

- Large volumes of structured, semi-structured, and unstructured data
- Agile sprints, quick iteration, and frequent code pushes
- Flexible, easy to use object-oriented programming
- Efficient, scale-out architecture instead of expensive, monolithic architecture

Classification

NoSQL database can be classified into four major categories (Han et al. 2011). The details are as follows:

Key-Value Stores Database: These systems typically store values and an index to find them, based on user-defined key. Examples: FoundationDB, DynamoDB, MemcacheDB, Redis, Riak, LevelDB, RocksDB, BerkeleyDB, Oracle NoSQL Database, GenieDB, BangDB, Chordless, Scalaris, KAI, FairCom c-tree, LSM, KitaroDB, upscaleDB, STSDB, Maxtable, RaptorDB, etc.

Document Stores Database: These systems usually store documents as just defined. The documents are indexed and a simple query mechanism is provided. Examples: Elastic, OrientDB, MongoDB, Cloud Datastore, Azure DocumentDB, Clusterpoint, CouchDB, Couchbase, MarkLogic, RethinkDB, SequoiaDB, RavenDB, JSON ODM, NeDB, Terrastore, AmisaDB, JasDB, SisoDB, DensoDB, SDB, iBoxDB, ThruDB, ReasonDB, IBM Cloudant, etc.

Graph Database: These systems database is designed for data whose relations are well represented as a graph. The kind of data could be social relations, public transport links, road maps, or network topologies. Examples: Neo4J, ArangoDB, Infinite Graph, Sparksee, TITAN, InfoGrid, HyperGraphDB, GraphBase, Trinity,

Bigdata, BrightstarDB, Onyx Database, VertexDB, FlockDB, Virtuoso, Stardog, Allegro, Weaver, Fallen 8, etc.

Column Database: These systems store extensible records that can be partitioned vertically and horizontally across nodes. Examples: Hadoop/HBase, Cassandra, Hortonworks, Scylla, HPCC, Accumulo, Hypertable, Amazon SimpleDB, Cloudata, MonetDB, Apache Flink, IBM Informix, Splice Machine, eXtremeDB Financial Edition, ConcourseDB, Druid, KUDU, Elassandra, etc.

All of the described databases are multimodal databases that is designed to support multiple data models against a single and integrated back end. Some examples are Datomic, GunDB, CortexDB, AlchemyDB, WonderDB, RockallDB, and FoundationDB.

NoSQL Database: Traditional databases are primarily relational, but in the NoSQL database fields, there are some new type of databases. Each type of database with an example is described as follows:

Key-Value Databases: Key-value (KV) stores use the associative array (also known as a map or dictionary) as their fundamental data model. In this model, data is represented as a collection of key-value pairs, such that each possible key appears at most once in the collection.

Redis

Redis is a key-value memory database: when Redis runs, data were entirely loaded into memory, so all the operations were run in memory and then periodically save the data asynchronously to the hard disk. The characteristics of pure memory operation make it to have very good performance; it can handle more than 100,000 read or write operation per second; (1) Redis supports list and set and various related operations; (2) the maximum of value limit to 1GB; and (3) the main drawback is that capacity of the database is limited by physical memory, so Redis cannot be used as Big Data storage, and scalability is poor.

Column-Oriented Databases

Though column-oriented database has not undermined the traditional store by row, but in architecture with data compression, hugely parallel processing, shared nothing, column-oriented database can main high performance of data analysis and business intelligence processing. Column-oriented databases are HBase, HadoopDB, Cassandra, Hypertable, Bigtable, PNUTS, etc.

Cassandra: Cassandra is an open-source database of Facebook. Its features are (1) the schema is very flexible and does not require to design database schema at first, and add or delete field is very convenient; (2) it supports range of queries, and (3) high scalability: a single point failure does not affect the whole cluster, and it supports linear expansion. Cassandra system is a distributed database system which was made of lots of database nodes; a write operation will be replicated to other nodes, and read request will be routed to a certain node. For a Cassandra cluster, only to add node can achieve the goal of scalability. In addition, Cassandra also supports rich data structure and powerful query language.

Document-Oriented Database

A **document-oriented database** is one type of NoSQL database designed for storing, retrieving, and managing document-oriented information, also known as **semi-structured data**. In contrast to relational databases and their notions of “relations” (or “tables”), these systems are designed around an abstract notion of a “document.” Unlike the key-value stores, these systems generally support secondary indexes and multiple types of documents (objects) per database and nested documents or lists. Like other NoSQL systems, the document stores do not provide ACID transactional properties.

MongoDB

MongoDB is an open-source database used by companies of all sizes, across all industries and for a wide variety of applications. It is an agile database that allows schemas to change quickly as

applications evolve, while still providing what the functionality developers expect from traditional databases, such as secondary indexes, a full query language, and strict consistency. MongoDB is built for scalability, performance and high availability, and scaling from single server deployments to large, complex multisite architectures. By leveraging in-memory computing, MongoDB provides high performance for both reads and writes. MongoDB's native replication and automated failover enable enterprise-grade reliability and operational flexibility.

Graph Database

A graph database is a database that uses graph structures with nodes, edges, and properties to represent and store data. A graph database is any storage system that provides index-free adjacency (Cattell 2011). This means that every element contains a direct pointer to its adjacent elements and no index lookups are necessary.

Neo4j

Neo4j is an [open-source](#) graph database, implemented in Java. The developers describe Neo4j as “embedded, disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables.” Neo4j is the most popular graph database.

NoSQL, a relatively new technology, has already attracted a significant amount of attention due to its use by massive websites like Amazon, Yahoo, and Facebook (Moniruzzaman and Hossain 2013).

NoSQL began within the domain of open source and a few small vendors, but continued growth in data and NoSQL has attracted many new players into the market. NoSQL solutions are attractive because they can handle huge quantities of data, relatively quickly, across a cluster of commodity servers that share resources. Additionally, most NoSQL solutions are open source, which gives them a price advantage over conventional commercial databases.

NoSQL Pros and Cons

Advantages

The major advantages of NoSQL database are described below.

Open Source

Most of the NoSQL databases are open source, offering development in the whole software market. According to Basho chief technology officer Justin Sheehy, open-source environment is healthy for NoSQL database, and it lets user perform technical evaluation at low cost. SQL (relational) versus NoSQL scalability is a controversial topic. Here is some more background to support this position. If new relational systems can do everything that a NoSQL system can, with analogous performance and scalability, and with the convenience of transactions and SQL, why would you choose a NoSQL system? Relational DBMSs have taken and retained majority market share over other competitors in the past 30 years: network, object, and XML DBMSs. Fruitful relational DBMSs have been built to handle other specific application loads in the past: read-only or read-mostly data warehousing, distributed databases, and now horizontally scaled databases.

Fast Data Processing

NoSQL databases usually process data faster than relational databases. Relational databases are mostly used by business transactions that require great precision. They thus generally subject all data to the same set of atomicity, consistency, isolation, and durability (ACID) fetters. Uppsala University professor Tore Risch explained about ACID. Atomicity means an update is performed completely or not at all. Consistency means no part of a transaction is allowed to break a database's rules. Isolation means each application runs transactions independently of other applications operating concurrently, and durability means that completed transactions will persist. Having to perform these restraints makes the relational database slower.

Scalability

The NoSQL approach presents huge advantages over SQL databases because it allows one to scale

an application to new levels. The new data services are based on truly scalable structures and architectures, built for the cloud and built for distribution, and are very attractive to the application developer. There is no need for DBA, no need for complicated SQL queries, and it is fast.

Disadvantages

There are some disadvantages to the NoSQL approach. Those are less visible at the developer level but are highly visible at the system, architecture, and operational levels.

Lack of skilled authority at the system level:

Not having a skilled authority to design a single, well-defined data model, regardless of the technology used, has its drawbacks. The data model may suffer from duplication of data objects (non-normalized model). This can happen due to the different object model used by different developers and their mapping to the persistency model. At the system level, one must also understand the limitations of the chosen data service, whether it is size, ops per second, concurrency model, etc.

Lack of interfaces and interoperability at the architecture level: Interfaces for the NoSQL data services are yet to be standardized. Even DHT, which is one of the simpler interfaces, still has no standard semantics, which includes transactions, non-blocking API, etc. Each DHT service used comes with its own set of interfaces. Another big issue is how different data structures, such as DHT and a binary tree, just as an example, share data objects. There are no intrinsic semantics for pointers in all those services. In fact, there is usually not even strong typing in these services – it is the developer's responsibility to deal with that. Interoperability is an important point, especially when data needs to be accessed by multiple services. A simple example: if Back office works in Java, and web serving works in PHP, can the data be accessed easily from both domains? Clearly one can use web services in front of the data as a data access layer, but that complicates things even more and reduces business agility, flexibility, and performance while increasing development overhead.

Less complaint to the operational realm:

The operational environment requires a set of tools that is not only scalable but also manageable and stable, be it on the cloud or on a fixed set of servers. When something goes wrong, it should not require going through the whole chain and up to the developer level to diagnose the problem. In fact, that is exactly what operation managers regard as an operational nightmare. Operation needs to be systematic and self-contained. With the current NoSQL services available in the market, this is not easy to achieve, even in managed environments such as Amazon.

Conclusion

NoSQL is a big and expanding area, covering classification of different types of NoSQL databases, performance measurement, advantages and disadvantages of NoSQL databases, and current state of adoption of NoSQL databases. This article provides an independent understanding of the strengths and weaknesses of various NoSQL database approaches to supporting applications that process huge volumes of data, as well as to provide a global overview of this non-relational NoSQL databases.

Further Reading

- Cattell, R. (2011). Scalable SQL and NoSQL data stores. *Acm Sigmod Record*, 39(4), 12–27.
- Stonebraker, M. (2010). SQL databases v. NoSQL databases. *Communications of the ACM*, 53(4), 10–11.
- Han, J., Haihong, E., Le, G., Du, J. (2011), October. Survey on NoSQL database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on* (pp. 363–366). IEEE.
- Moniruzzaman, A. B. M., & Hossain, S. A. (2013). NoSQL database: New era of databases for big data analytics – Classification, characteristics and comparison. *International Journal of Database Theory and Application*, 6(4), 1–14.

NSF

► [Big Data Research and Development Initiative \(Federal, U.S.\)](#)

Nutrition

Qinghua Yang¹ and Yixin Chen²

¹Department of Communication Studies, Texas Christian University, Fort Worth, TX, USA

²Department of Communication Studies, Sam Houston State University, Huntsville, TX, USA

Nutrition is a science that helps people to make good choices of foods to keep healthy, by identifying the amount of nutrients they need and the amount of nutrients each food contains. Nutrients are chemicals obtained from diet and are indispensable to people's health. Keeping a balanced diet containing all essential nutrients can prevent people from diseases caused by nutritional deficiencies such as scurvy and pellagra.

Although the United States has one of the most advanced nutrition sciences in the world, the nutrition status of the U.S. population is not optimistic. While nutritional deficiencies as a result of dietary inadequacies are not very common, many Americans are suffering from overconsumption-related diseases. Due to the excessive intake of sugar and fat, the prevalence of overweight and obesity in the American adult population increased from 47% to over 65% over the past three decades, currently with two-thirds of American adults being overweight and among whom 36% being obese. Overweight and obesity are concerns not only for the adult population, but also for the childhood population, with one third of American children being overweight or obese. Obesity kills more than 2.8 million Americans every year, and the obesity-related health problems cost American taxpayers more than \$147 billion every year. Thus, reducing the obesity prevalence in the United States has become a national health priority.

Big data research on nutrition holds tremendous promise for preventing obesity and improving population health. Recently, researchers have been trying to apply big data to nutritional research, by taking advantages of the increasing amount of nutritional data and the accumulation of nutritional studies. Big data is a collection of

data sets, which are large in volume and complex in structure. For instance, the data managed by America's leading health care provider Kaiser is more than 4,000 times the amount of information stored in the Library of Congress. As to data structure, nutritional data and ingredients are really difficult to normalize. The volume and complexity of nutritional big data make it difficult to process them using traditional data analytic techniques.

Big data analyses can provide more valuable information than traditional data sets and reveal hidden patterns among variables. In a big data study sponsored by the National Bureau of Economic Research, economists Matthew Harding and Michael Lovenheim analyzed data of over 123 million purchasing decisions on food and beverage made in the U.S. between 2002 and 2007 and simulated the effects of various taxes on Americans' buying habits. Their model predicted that an increase of 20% tax on sugar would reduce Americans' total caloric intake by 18% and reduce sugar consumption by over 16%. Based on their findings, they proposed a new policy of implementing a broad-based tax on sugar to improve public health. In another big-data study on human nutrition, two researchers at West Virginia University tried to understand and monitor the nutrition status of a population. They designed intelligent data collection strategies and examined the effects of food availability on obesity occurrence. They concluded that modifying environmental factors (e.g., availability of healthy food) could be the key in obesity prevention.

Big data can be applied to self-tracking, that is, monitoring one's nutrition status. An emerging trend in big data studies is quantified self (QS), which refers to keeping track of one's nutritional, biological and physical information, such as calories consumed, glycemic index, and specific ingredients of food intake. By pairing the self-tracking device with a web interface, the QS solutions can provide users with nutrient-data aggregation, infographic visualization, and personal recommendations for diet.

Big data can also enable researchers to monitor the global food consumption. One pioneering project is the Global Food Monitoring Group

conducted by the George Institute for global health with participations from 26 countries. With the support of these countries, the Group is able to monitor the nutrition composition of various foods consumed around the world, identify the most effective food reformulation strategies, and explore effective approaches on food production and distribution by food companies in different countries.

Thanks to the development of modern data collection and analytic technologies, the amount of nutritional, dietary, and biochemical data continues to increase at a rapid pace, along with a growing accumulation of nutritional epidemiologic studies during this time. The field of nutritional epidemiology has witnessed a substantial increase in systematic reviews and meta-analyses over the past two decades. There were 523 meta-analyses and systematic reviews within the field of nutritional epidemiology in 2013 versus just 1 in 1985. However, in the era of “big data”, there is an urgent need to translate big-data nutrition research to practice, so that doctors and policymakers can utilize this knowledge to improve individual and population health.

Controversy

Despite the exciting progress of big-data application in nutrition research, several challenges are equally noteworthy. First, to conduct big-data nutrition research, researchers often need access to a complete inventory of foods purchased in all retail outlets. This type of data, however, is not readily available and gathering such information site by site is a time-consuming and complicated process. Second, information provided by nutrition big data may be incomplete or incorrect. For example, when doing self-tracking for nutrition

status, many people fail to do consistent daily documentation or suffer from poor recall of food intake. Also, big data analyses may be subject to systematic biases and generate misleading research findings. Lastly, since an increasing amount of personal data is being generated through quantified self-tracking devices, it is important to consider privacy rights in personal data. That individuals’ personal nutritional data should be well-protected and that data shared and posted publicly should be used appropriately are key ethical issues for nutrition researchers and practitioners. In light of these challenges, technical, methodological, and educational interventions are needed to deal with issues related to big-data accessibility, errors and abuses.

Cross-References

- [Biomedical Data](#)
- [Data Mining](#)
- [Health Informatics](#)

Further Reading

- Harding, M., & Lovenheim, M. (2017). The effect of prices on nutrition: Comparing the impact of product- and nutrient-specific taxes. *Journal of Health Economics*, 53.
- Insel, P., et al. (2013). *Nutrition*. Boston: Jones and Bartlett Publishers.
- Satija, A., & Hu, F. (2014). Big data and systematic reviews in nutritional epidemiology. *Nutrition Reviews*, 72(12).
- Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2).
- WVU Today. WVU researchers work to track nutritional habits using ‘Big Data’. <http://wvutoday.wvu.edu/n/2013/01/11/wvu-researchers-workto-track-nutritional-habits-using-big-data>. Accessed Dec 2014.

Online Advertising

Yulia A. Strekalova
College of Journalism and Communications,
University of Florida, Gainesville, FL, USA

In a broad sense, online advertising means advertising through cross-referencing on a business's own web portal or on the websites of other online businesses. The goal of online advertising is to attract attention to advertised websites and products and, potentially, lead to an enquiry about a project, mail list subscription, or product purchase. Online advertising creates new cost-saving opportunities for businesses by reducing some of the risks of ineffective advertising resources. Online advertising types include banners, targeted ads, and social media community interactions, and each type requires careful planning and consideration of potential ethical challenges.

Online advertising analytics and measurement is necessary to assess the effectiveness of advertising efforts and the return on the investment of funds. However, measurement is challenged by the fact that advertising across media platforms is increasingly interactive. For example, a TV commercial may lead to an online search, which will result in a relevant online ad, which may lead to a sale. The vast amounts of data and powerful analytics are necessary to allow advertisers performing high-definition cross-channel analyses of the public and its behaviors, evaluate the

return on investments across media, generate predictive models, and modify their campaigns in near-real time. The proliferation of data collection gave rise to increased concerns among the Internet users and advocacy groups. As the user data are collected by shared among multiple parties, they may amount to become personally identifiable to a particular person.

Types of Online Advertising

Online advertising, a multibillion-dollar industry today, started from a single marketing email offering a new computer system sent in 1978 to 400 users of the Advanced Research Projects Agency Network (ARPAnet). While the reactions to this first online advertising campaign were negative and identified the message as spam, email and forum-based advertising continued to develop and grow. In 1993, a company called Global Network Navigator sold the first clickable online ad. AT&T, one of the early adopters of this advertising innovation, received clicks from almost half of the Internet users who were exposed to its "Have you ever clicked your mouse right HERE? – You will." banner ad. In 1990s, online advertising industry was largely fragmented, but first ad networks started to appear and offer their customers opportunities to develop advertising campaigns that will place ads across a diverse set of websites and reach particular audience segments. An advertising banner may be placed on

high-traffic sites statically for a predefined period of time. While this method may be the least costly and targeted to a niche audience, it does not allow for rich data collection. Banner advertising is a less sophisticated form of online advertising. Banner advertising could also be used as a hybrid of cost per mille (CPM), or cost per thousand, as another advertising option which will deliver an ad to website users. This option is usually priced in a multiple of 1,000 impressions (or the number of times an ad was shown) and an additional cost for clicks. It also allows businesses to assess how many times an ad was shown. However, this method is limited in its ability to measure if the return on an investment in advertising covered the costs. However, proliferation of banners on sites and the overall volume of information on sites lead to “banner blindness” among the Internet users. In addition, with rapid increase of mobile phones as Internet connection devices, the average effectiveness of banners became even lower. The use of banner and pop-up ads increased in the late 1990s and early 2000s, but the users of the Internet started to block these ads with pop-up blockers, and the clicks on banner ads dropped to about 0.1%.

The next innovation in the online advertising is tied to the growth in sophistication of search engines. The search engines started to allow advertisers to place ad relevant to particular keywords. Tying advertising to relevant search keywords gave rise to the pay-per-click (PPC) advertising. PPC provides advertisers with most robust data to assess if expended costs generated sufficient return. PPC advertising means that advertisers are charged per click on an ad. This advertising method ties exposure to advertising to an action from a potential consumer thus providing advertisers with the data on the sites that are more effective. Google AdWords is an example of pay-per-click advertising, which is linked to the keywords and phrases used in search. AdWords ads are correlated with these keywords and shown only to the Internet users with relevant searches. By using PPC in conjunction with a search engine, like Google, Bing, or Yahoo, advertisers can also

obtain insights on the environment or search terms that led a consumer to the ad in the first place.

Online advertising may also include direct newsletter advertising delivered to potential customers who have purchased before. However, the decision to use this way of advertising should be coupled with an ethical way of employing it. Email addresses became a commodity and can be bought. However, a newsletter sent to users who never bought from a company may fire back and lead to unintended negative consequences. Overall, this low-cost advertising method can be effective in keeping past customers informed about new products and other campaigns run by the company.

Social media is another advertising channel, which is rapidly growing in its popularity. Social media networks created repositories of psychographic data, which include user-reported demographic information, hobbies, travel destinations, lifetime events, and topics of interest. Social media can be used as more traditional advertising channels for PPC ad placements. However, they can also serve as a base for customer engagement. Social media, although require a commitment and time investment from advertisers, may generate brand loyalty. Social media efforts, therefore, require careful evaluation as they can be both costly in terms of direct advertising costs and the cost of time spent by company employees on developing and executing social media campaign and keeping the flow of communication active. Data collected from social media channels can be analyzed on the individual level, which was nearly impossible with earlier online advertising methods. Companies can collect information about specific user communication and engagement behavior, track communication activities of individual users, and analyze comments shared by the social media users. At the same time, aggregate data may allow for general sentiment analysis to assess if overall comments about a brand are positive or negative and seek out product-related signals shared by users. Social media evaluation, however, is challenged by the absence of deep understanding of the audience engagement

metrics and lack of industry-wide benchmarks and evaluation standards. As a fairly new area of advertising, social media evaluation of likes, comments, and shares may be interpreted in a number of ways. Social media networks provide a framework for a new type of advertising, community exchange, but they also are channels of online advertising through real-time advertising targeting. It is likely that focused targeting will continue to be the focus of advertisers as it leads to the increases in the effectiveness of advertising efforts. At the same time, tracking of user web behavior throughout the Web creates privacy concerns and policy challenges.

Targeting

Innovations in online advertising introduced targeting techniques that based advertising on the past browsing and purchase behaviors of Internet users. Proliferation of data collection enabled advertisers to target potential clients based on a multitude of web activities, like site browsing, key word searchers, past purchasing across different merchants, etc. These targeting techniques led to the development of data collection systems that track user activity in real time and make decisions to advertise or not advertise right as the user is browsing a particular page. Online advertising lacks rigorous standardization and several recent targeting typologies have been proposed. Reviewing strategies for online advertising, Gabriela Taylor identifies nine distinct targeting methods, which overlap or complement the discussion of targeting methods proposed by other authors. In general, targeting refers to situation when ads that are shown to an Internet user are relevant to their interests. The latter are determined by the keywords used on searchers, pages visited, or online purchases made.

Contextual targeting ads are delivered to web users based on the content of the sites these users visit. In other words, contextually targeted advertising matches ads to the content of the webpage an Internet user is browsing. Systems managing contextual advertising scan websites for

keywords and place ads that match these keywords most closely. For example, a user viewing a website about gardening may see ads for gardening and house-keeping magazines or home improvement stores.

Geo, or local, targeting is focused on the determination of the geographical location of a website visitor. This information, in turn, is used to deliver ads that are specific to a particular location, country, region or state, city, or metro area. In some cases, targeting can go as deep as an organizational level. Internet protocol (IP) address, assigned to each device participating a computer network, is used as the primary data point in this targeting method. The use of this method may prevent the delivery of ads to users where product or service is not available – for example, a content restriction for Internet television or region-specific advertising that complies with regional regulations.

Demographic targeting, as implied by its name, tailors ads based on website users' demographic information, like gender, age, income and education level, marital status, ethnicity, language preferences, and other data points. Users may supply this information is social networking site registration. The sites, additionally, may also encourage its users to “complete” their profiles after the initial registration to get access to the fullest set of data.

Behavioral targeting looks at users' declared or expressed interests to tailor the content of delivered ads. Web-browsing information, data on the pages visited, the amount of time spent on particular pages, meta-data for the links that were clicked, the searches conducted recently, and information about recent purchases is collected and analyzed by advertisement delivery systems to select and display the most relevant ads. In a sense, website publishers can create user profiles based on the collected data and use it to predict future browsing behavior and potential products of interest. This approach, using rich past data, allows advertisers to target their ads more effectively to the page visitors who are more likely to have interest in these products or services. Combined with other strategies, including contextual,

geographic, and demographic targeting, this approach may lead to finely tuned and interest-tailored ads. The approach proves effective as several studies showed that also Internet users prefer to have no ads on the web-pages they visit, they favor relevant ads over random ones.

DayPart and time-based targeting is run during specific times of the day or the week, for example, 10 am to 10 pm local time Monday through Friday. Ads targeted based on this method are displayed only during these days and times and go off during the off-times. Ads run through DayPart campaigns may focus on time-limited offers and create a sense of urgency among audience members. At the same time, such ads may create an increased sense of monitoring and lack of privacy among the users exposed to these ads.

Real-time targeting allows for the ad placement systems to place bids for advertisement placement in real time. Additionally, this advertising method allows to track every unique site user and collect real-time data to assess the likelihood of each visitor to make a purchase.

Affinity targeting creates a partnership between a product producer and an interest-based organization to promote the use of a third-party product. This method targets customers who share interest in a particular topic. These customers are assumed to have positive attitude toward a website they visit and therefore have a positive attitude toward more relevant advertising. This method is akin to niche advertising, and its success is based on the close match between the advertising content and that of the passions and interests of website users.

Look-alike targeting aims to identify prospective customers who are similar to the advertiser's customer base. Original customer profiles are determined based on the website use and previous behaviors of active customers. These profiles are then matched against a pool of independent Internet users who share common attributes and behaviors and are the likely targets for an advertised product. The challenge with identifying these look-alike audiences is challenged by the large number of possible input data points which may

or may not be defining for a particular behavior or user group.

Act-alike targeting is an outcome of predictive analytics. Advertisers using this method define profiles of customers based on their information consumption and spending habits. Customers and their past behaviors are identified; they are segmented into groups to predict their future purchase behavior. The goal of this method is to identify the most loyal group of customers, who generate revenue for the company and engage with this group in a most effective and supportive way.

Privacy Concerns

Technology is developing at a speed too rapid for policy-making to catch up. Whichever advertising targeting method is used, each is based on an extended collections and analysis of personal and behavioral data for each user. Ongoing and potentially pervasive data collection raises important privacy questions and concerns. Omer Tene and Jules Polonetsky identify several privacy risks associated with big data. First is an incremental adverse effect on privacy from an ongoing accumulation of information. More and more data points are collected about individual Internet users and once information about real identify has been linked to a virtual identify of a user, the anonymity is lost. Furthermore, disassociation of a user with a particular service may be insufficient to break a previously existing link as other networks and online resources may have already harvested missing data points. Second area of privacy risks is an automated decision-making process. These automated algorithms may lead to discrimination and self-determination. Targeting and profiling used in online advertising gives ground to potential threats to the free access to information and open, democratic society. The third area of privacy concerns is predictive analysis, which may identify and predict stigmatizing behaviors or characteristics, like susceptibility to disease or undisclosed sexual orientation. In addition, predictive analysis may give ground to social

stratification by putting users in like-behaving clusters and ignoring outliers and minority groups. Finally, the fourth area of concern is the lack of access to information and exclusion of smaller organizations and individuals from the access to the benefits of big data. Large organizations are able to collect and use big data to price products close to an individual's reservation price or cornering an individual with a deal impossible to resist. At the same time, large organizations are seldom forthcoming with sharing individuals' information with these individuals in an assessable and understandable format.

Cross-References

- [Content Management System \(CMS\)](#)
- [Data-Information-Knowledge-Wisdom \(DIKW\) Pyramid, Framework, Continuum](#)
- [Predictive Analytics](#)
- [Social Media](#)

Further Reading

- Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Hoboken: Wiley.
- Taylor, G. (2013). *Advertising in a digital age: Best practices & tips for paid search and social media advertising*. Global & Digital.
- Tene, O., & Polonetsky, J. (2013). Privacy in the age of big data: A time for big decisions. *Stanford Law Review Online*, 11/5.
- Turow, J. (2012). *The daily you: How the advertising industry is defining your identity and your worth*. New Haven: Yale University Press.

Online Analytical Processing

- [Data Mining](#)

Online Commerce

- [E-Commerce](#)

Online Identity

Catalina L. Toma

Communication Science, University of Wisconsin-Madison, Madison, WI, USA

Identity refers to the stable ways in which individuals or organizations think of and express themselves. The availability of big data has enabled researchers to examine online communicators' identity using generalizable samples. Empirical research to date has focused on personal, rather than organizational, identity, and on social media platforms, particularly Facebook and Twitter, given that these platforms require users to present themselves and their daily reflections to audiences. Research to date has investigated the following aspects of online identity: (1) *expression*, or how users express who they are, especially their personality traits and demographics (e.g., gender, age) through social media activity; (2) *censorship*, or how users suppress their urges to reveal aspects of themselves on social media; (3) *detection*, or the extent to which it is possible to use computational tools to infer users' identity from their social media activity; (4) *audiences*, or who users believe accesses their social media postings and whether these beliefs are accurate; (5) *families*, or the extent to which users include family ties as part of their identity portrayals; and (6) *culture*, or how users express their identities in culturally determined ways. Each of these areas of research is described in detail below.

Identity Expression

In its early days, the Internet appealed to many users because it allowed them to engage with one another anonymously. However, in recent years, users have overwhelmingly migrated toward personalized interaction environments, where they reveal their real identities and often connect with members of their offline networks. Such is the case with social media platforms. Therefore,

research has taken great interest in how users communicate various aspects of their identities to their audiences in these personalized environments.

One important aspect of people's identities is their personality. Big data has been used to examine how personality traits get reflected in people's social media activity. How do people possessing various personality traits talk, connect, and present themselves online? The development of the myPersonality Facebook application was instrumental in addressing these questions. myPersonality administers personality questionnaires to Facebook users and then informs them of their personality typology in exchange for access to all their Facebook data. The application has attracted millions of volunteers on Facebook and has enabled researchers to correlate Facebook activities with personality traits. The application, used in all the studies summarized below, measures personality using the Big Five Model, which specifies five basic personality traits: (1) extraversion, or an individual's tendency to be outgoing, talkative, and socially active; (2) agreeableness, or an individual's tendency to be compassionate, cooperative, trusting, and focused on maintaining positive social relations; (3) openness to experience, or an individual's tendency to be curious, imaginative, and interested in new experiences and ideas; (4) conscientiousness, or an individual's tendency to be organized, reliable, consistent, and focused on long-term goals and achievement; and (5) neuroticism, or an individual's tendency to experience negative emotions, stress, and mood swings.

One study conducted by Yoram Bachrach and his colleagues investigated the relationship between Big Five personality traits and Facebook activity for a sample of 180,000 users. Results show that individuals high in extraversion had more friends, posted more status updates, participated in more groups, and "liked" more pages on Facebook; individuals high in agreeableness appeared in more photographs with other Facebook users but "liked" fewer Facebook pages; individuals high in openness to experience posted more status updates, participated in more groups, and "liked" more Facebook pages;

individuals high in conscientiousness posted more photographs but participated in fewer groups and "liked" fewer Facebook pages; and individuals high in neuroticism had fewer friends but participated in more groups and "liked" more Facebook pages. A related study, conducted by Michal Kosinski and his colleagues, replicated these findings on a sample of 350,000 American Facebook users, the largest dataset to date on the relationship between personality and Internet behavior.

Another study examined the relationship between personality traits and word usage in the status updates of over 69,000 English-speaking Facebook users. Results show that personality traits were indeed reflected in natural word use. For instance, extroverted users used words reflecting their sociable nature, such as "party," whereas introverted users used words reflecting their more solitary interests, such as "reading" and "Internet." Similarly, highly conscientious users expressed their achievement orientation through words such as "success," "busy," and "work," whereas users high in openness to experience expressed their artistic and intellectual pursuits through words like "dreams," "universe," and "music."

In sum, this body of work shows that people's identity, operationalized as personality traits, is illustrated in the actions they undertake and words they use on Facebook. Given social media platforms' controllable nature, which allows users time to ponder their claims and the ability to edit them, researchers argue that these digital traces likely illustrate users' *intentional* efforts to communicate their identity to their audience, rather than being unintentionally produced.

Identity Censorship

While identity expression is frequent in social media and, as discussed above, illustrated by behavioral traces, sometimes users suppress identity claims despite their initial impulse to divulge them. This process, labeled "last-minute self-censorship," was investigated by Sauvik Das and Adam Kramer using data from 3.9 million

Facebook users over a period of 17 days. Censorship was measured as instances when users entered text in the status update or comment boxes on Facebook but did not post it in the next 10 min. The results show that 71% of the participants censored at least one post or comment during the time frame of the study. On average, participants censored 4.52 posts and 3.20 comments. Notably, 33% of all posts and 13% of all comments written by the sample were censored, indicating that self-censorship is a fairly prevalent phenomenon. Men censored more than women, presumably because they are less comfortable with self-disclosure. This study suggests that Facebook users take advantage of controllable media affordances, such as editability and unlimited composition time, in order to manage their identity claims. These self-regulatory efforts are perhaps a response to the challenging nature of addressing large and diverse audiences, whose interpretation of the poster's identity claims may be difficult to predict.

Identity Detection

Given that users leave digital traces of their personal characteristics on social media platforms, research has been concerned with whether it is possible to *infer* these characteristics from social media activity. For instance, can we deduce users' gender, sexual orientation, or personality from their explicit statements and patterns of activity? Is their identity implicit in their social media activity, even though they might not disclose it explicitly?

One well-publicized study by Michal Kosinski and his colleagues sought to predict Facebook users' personal characteristics from their "likes" – that is, Facebook pages dedicated to products, sports, music, books, restaurant, and interests – that users can endorse and with which they can associate by clicking the "like" button. The study used a sample of 58,000 volunteers recruited through the myPersonality application. Results show that, based on Facebook "likes," it is possible to predict a user's ethnic identity (African-American vs. Caucasian) with 95% accuracy,

gender with 93% accuracy, religion (Christian vs. Muslim) with 82% accuracy, political orientation (Democrat vs. Republican) with 85% accuracy, sexual orientation among men with 88% accuracy and among women with 75% accuracy, and relationship status with 65% accuracy. Certain "likes" stood out as having particularly high predictive ability for Facebook users' personal characteristics. For instance, the best predictors of high intelligence were "The Colbert Report," "Science," and, unexpectedly, "curly fries." Conversely, low intelligence was indicated by "Sephora," "I Love Being a Mom," "Harley Davidson," and "Lady Antebellum."

In the area of personality, two studies found that users' extraversion can be most accurately inferred from Facebook profile activity (e.g., group membership, number of friends, number of status updates); neuroticism, conscientiousness, and openness to experience can be reasonably inferred; and agreeableness cannot be inferred at all. In other words, Facebook activity renders extraversion highly visible and agreeableness opaque.

Language can also be used to predict online communicators' identity, as shown by Andrew Schwartz and his colleagues in a study of 15.4 million Facebook status updates, totaling over 700 million words. Language choice, including words, phrases, and topics of conversation, was used to predict users' gender, age, and Big Five personality traits with high accuracy.

In sum, this body of research suggests that it is possible to infer many facets of Facebook users' identity through automated analysis of their online activity, regardless of whether they explicitly choose to divulge this identity. While users typically choose to reveal their gender and ethnicity, they can be more reticent in disclosing their relational status or sexual orientation and might themselves be unaware of their personality traits or intelligence quotient. This line of research raises important questions about users' privacy and the extent to which this information, once automatically extracted from Facebook activity, should be used by corporations for marketing or product optimization purposes.

Real and Imagined Audience for Identity Claims

The purpose of many online identity claims is to communicate a desired image to an audience. Therefore, the process of identity construction involves understanding the audience and targeting messages to them. Social media, such as Facebook and Twitter, where identity claims are posted very frequently, pose a conundrum in this regard, because audiences tend to be unprecedentedly large, sometimes reaching hundreds and thousands of members, and diverse. Indeed, “friends” and “followers” are accrued over time and often belong to different social circles (e.g., high school, college, employment). How do users conceptualize their audiences on social media platforms? Are users’ mental models of their audiences accurate?

These questions were addressed by Michael Bernstein and his colleagues in a study focusing specifically on Facebook users. The study used a survey methodology, where Facebook users indicated their beliefs about how many of their “friends” viewed their Facebook postings, coupled with large-scale log data for 220,000 Facebook users, where researchers captured the actual number of “friends” who viewed users’ postings. Results show that, by and large, Facebook users underestimated their audiences. First, they believed that any specific status update they posted was viewed, on average, by 20 “friends,” when in fact it was viewed by 78 “friends.” The median estimate for the audience size for any specific post was only 27% of the actual audience size, meaning that participants underestimated the size of their audience by a factor of 4. Second, when asked how many total audience members they had for their profile postings during the past month, Facebook users believed it was 50, when in fact it was 180. The median perceived audience for the Facebook profile, in general, was only 32% of the actual audience, indicating that users underestimated their cumulative audience by a factor of 3. Slightly less than half of Facebook users indicated they wanted a larger audience for their identity claims than they thought they had, ironically failing to

understand that they did in fact have this larger audience. About half of Facebook users indicated that they were satisfied with the audience they thought they had, even though their audience was actually much greater than they perceived it to be. Overall, this study highlights a substantial mismatch between users’ beliefs about their audiences and their actual audiences, suggesting that social media environments are translucent, rather than transparent, when it comes to audiences. That is, actual audiences are somewhat opaque to users, who as a result may fail to properly target their identity claims to their audiences.

Family Identity

One critical aspect of personal identity is family ties. To what extent do social media users reveal their family connections to their audience, and how do family members publically talk to one another on these platforms? Moira Burke and her colleagues addressed these questions in the context of parent-child interactions on Facebook. Results show that 37.1% of English-speaking US Facebook users specified either a parent or child relationship on the site. About 40% of teenagers specified at least one parent on their profile, and almost half of users age 50 or above specified a child on their profile. The most common family ties were between mothers and daughters (41.4% of all parent-child ties), followed by mothers and sons (26.8%), fathers and daughters (18.9%), and least of all fathers and sons (13.1%). However, Facebook communication between parents and children was limited, accounting for only 1–4% of users’ public Facebook postings. When communication did happen, it illustrated family identities: Parents gave advice to children, expressed affection, and referenced extended family members, particularly grandchildren.

Cultural Identity

Another critical aspect of personal identity is cultural identity. Is online communicators’ cultural

identity revealed by their communication patterns? Jaram Park and colleagues show that Twitter users create emoticons that reflect an individualistic or collectivistic cultural orientation. Specifically, users from individualistic cultures preferred horizontal and mouth-oriented emoticons, such as :), whereas users from collectivistic cultures preferred vertical and eye-oriented emoticons, such as ^_^ . Similarly, a study of self-expression using a sample of four million Facebook users from several English-speaking countries (USA, Canada, UK, Australia) shows that members of these cultures can be differentiated through their use of formal or informal speech, the extent to which they discuss positive personal events, and the extent to which they discuss school. In sum, this research shows that cultural identity is evident in linguistic self-expression on social media platforms.

Cross-References

- ▶ [Anonymity](#)
- ▶ [Behavioral Analytics](#)
- ▶ [Facebook](#)
- ▶ [Privacy](#)
- ▶ [Profiling](#)
- ▶ [Psychology](#)

Further Reading

- Bachrach, Y., et al. (2012). Personality and patterns of Facebook usage. In *Proceedings of the 3rd Annual Web Science Conference* (pp. 24–32). Association for Computing Machinery.
- Bernstein, M., et al. (2013). Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 21–30). Association for Computing Machinery.
- Burke, M., et al. (2013). Families on Facebook. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)* (pp. 41–50). Association for the Advancement of Artificial Intelligence.
- Das, S., & Kramer, A. (2013). Self-censorship on Facebook. In *Proceedings of the 2013 Conference on Computer-Supported Cooperative Work* (pp. 793–802). Association for Computing Machinery.
- Kern, M., et al. (2014). The online social self: An open vocabulary approach to personality. *Assessment*, 21, 158–169.

Kosinski, M., et al. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110, 5802–5805.

Kramer, A., & Chung, C. (2011). Dimensions of self-expression in Facebook status updates. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)* (pp. 169–176). Association for the Advancement of Artificial Intelligence.

Park, J., et al. (2014). Cross-cultural comparison of non-verbal cues in emoticons on twitter: Evidence from big data analysis. *Journal of Communication*, 64, 333–354.

Schwartz, A., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, 8, e73791.

Ontologies

Anirudh Prabhu
Tetherless World Constellation, Rensselaer
Polytechnic Institute, Troy, NY, USA

Synonyms

[Computational ontology](#); [Knowledge graph](#);
[Semantic data model](#); [Taxonomy](#); [Vocabulary](#)

Definition

Ontology provides a rich description of the:

- Terminology, concepts, nomenclature
- Relationships among and between concepts and individuals
- Sentences distinguishing concepts, refining definitions and relationships (constraints, restrictions, regular expressions)

Relevant to a particular domain or area of interest (Kendall and McGuinness 2019).

An ontology defines a common vocabulary for researchers who need to exchange information in a domain. It can include machine-interpretable definitions of basic concepts in the domain and relations among them (Noy and McGuinness 2001).

The specification of an ontology can vary depending on the reason for developing an ontology, the domain of the ontology, and its intended use.

History

The first ever written use of the word ontology comes from the Latin word “*ontologia*” coined in 1613, independently by two philosophers, Rudolf Göckel in his “*Lexicon Philosophicum*” and Jacob Lorhard in his “*Theatrum Philosophicum*” (Smith and Welty 2001). In English, the first recorded use of the word was seen in Bailey’s dictionary of 1721, where ontology is defined as “an Account of being in the Abstract.” (Smith and Welty 2001).

Artificial intelligence researchers in 1980s were the first to adopt this word in computer science. These researchers recognized that one could create ontologies (information models) which could leverage automated reasoning capabilities.

In the 1990s, Thomas Gruber wrote two influential papers titled “Toward Principles for the Design of Ontologies Used for Knowledge Sharing” and “A Translation Approach to Portable Ontology Specifications”. In the first paper, he introduced the notion of ontologies as designed artifacts. He also provides a guide for designing formal ontologies based on five design criteria. These criteria will be described in more detail later in the entry.

In 2001, Tim Berners-Lee, Jim Hendler, and Ora Lassila described the evolution of the then existing Web (World Wide Web) to Semantic Web. In that article, Berners-Lee informally defines semantic web as “an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.” (Kendall and McGuinness 2019).

As the web of data grows, the amount of machine-readable data that can be used with ontologies also increases. And with the World Wide Web being highly ubiquitous, data from a

variety of disciplines (or domains) are thus increasingly available for discovery, access, and use.

Ontology Components

Most ontologies are structurally similar. Common components of ontologies include:

- *Individuals (or instances)* are the basic ground level components of an ontology. An ontology together with a set of individual instances is commonly called a *knowledge base*.
- *Classes* describe concepts in the domain. A class can have *subclasses* that represent concepts more specific than their superclass. Classes are the focus of most ontologies.
- *Attributes (or annotations)* are assigned to classes or subclasses to help describe classes in an ontology.
- *Relations (or property relations)* help specify how the entities in an ontology are related to other entities.
- *Restrictions* are formal conditions that affect the acceptance of assertions as input.

Ontology Engineering

Ontology engineering encompasses the study of the ontology development process including methodologies, tools, and languages for building ontologies.

Why Develop Ontologies?

There are many reasons to develop an ontology. For example, ontologies help share a common understanding of the structure of information among people or software agents. Ontologies are also developed to make domain assumptions explicit and to enable the analysis and reuse of domain knowledge (Noy and McGuinness 2001).

Ontology Development Process

There isn't a single "correct" way to develop an ontology. This entry will be following a combination of the methodologies described in Kendall and McGuinness (2019) and Noy and McGuinness (2001).

Development of an ontology usually starts from its use case(s). Most ontology development projects are driven by the research questions that are derived from a specific use case. These questions would typically be the starting point for developing the ontology.

Once the use case has been completed, it is time to determine the domain and scope of the ontology. The domain and scope of the ontology can be defined by answer several basic questions like (1) What is the domain that the ontology will cover? (2) For what are we going to use this ontology? (3) For what types of questions will the information in the ontology provide answers? (4) Who will use and maintain the ontology? (Noy and McGuinness 2001).

These answers may change and evolve of the time of the development process, but they help limit the scope of the ontology. Another way of determining the scope of the ontology is to use competency questions. Competency questions are ones that the users would want to use the ontology to answer. Competency questions can also be used for evaluating the success of the ontology later in the development process.

The next step in the ontology development process is to enumerate the terms required in the ontology. As seen in the flowchart, there are many sources to the list the terms. There is a plethora of the well-developed ontologies and vocabularies available for use. So, if parts (or all) of the other available ontologies fit into the ontology being developed, they should without a question be imported and used. Reusing existing ontologies and vocabularies is one of the best practices in ontology development. Some other sources include database schema, data dictionaries, text documents, and terms obtained from domain experts both within and outside the primary development team.

Once a final "term list" (or concept spreadsheet) has been compiled, it is time to build the ontology. This ontology can be built one of two ways. One method is to build the ontology by manually using domain specific concepts, related domain concepts, authoritative vocabularies, vetted definitions, and supporting citations from literature. The other method is to use automated tools and scripts to generate ontology from concept spreadsheets. Human evaluation of the results in the automated ontology generation process remains important to ensure the ontology generated accurately represent the expert's view of the domain.

After the ontology has been generated, it should be available to the users to explore or use in different applications. Using an interactive ontology browser enables discovery, review, and commentary of concepts.

Ontology development is an ongoing process and by no means is finished once the ontology has been generated. The ontology is maintained and curated by the ontology development team, domain collaborators, invited experts, and the consumers. The user community can also engage in commentary and recommend changes to the ontology.

Design Criteria for Ontologies

Numerous design decisions need to be made while developing an ontology. It is important to follow a guide or a set of objective criteria in order to make a "well designed" ontology. Generally, the design criteria suggested by Thomas Gruber are used for this purpose. Gruber proposed a preliminary set of criteria for ontologies whose purpose is knowledge sharing and interoperation among programs (and applications) based on a shared conceptualization (Gruber 1995).

Clarity: An ontology should be clear in communicating the meaning of its terms. Definitions should be objective and where possible definitions should be stated in logical axioms. "A *complete* definition (a predicate defined by

necessary and sufficient conditions) is preferred over a partial definition (defined by only necessary or sufficient conditions). All definitions should be documented with natural language” (Gruber 1995).

Coherence: All the inferences in an ontology should be logically consistent with its definitions (both formal and informal) of its concepts. “If a sentence that can be inferred from the axioms contradicts a definition or example, then the ontology is incoherent” (Gruber 1995).

Extendibility: The ontology should be designed by taking potential future extensions into account. When a user extends the ontology in the future, there should not be a need to revise existing definitions.

Minimal encoding bias: An ontology should be independent of the issues of the implementing language. An encoding bias occurs when the developers make design choices based on ease of implementation.

Minimal ontological commitments: An ontology should make very few claims about the domain being modeled. This gives the user the freedom to specialize or generalize the ontology depending on their need.

Ontology Languages

An ontology language is a formal language used to encode the ontology. Some of the commonly used ontology languages are RDF+RDF(S), OWL, SKOS, KIF, and DAML+OIL (Kalibatiene and Vasilecas 2011).

RDF + RDFS: “The Resource Description Framework (RDF) is a recommendation for describing resources on the web developed by the World Wide Web Consortium (W3C). It is designed to be read and understood by computers, not displayed to people” (Kalibatiene and Vasilecas 2011).

“An RDF Schema (RDFS) is a RDF vocabulary that provides identification of classes, inheritance relations for classes, inheritance relations

for properties and, domain and range properties” (Kendall and McGuinness 2019).

OWL: The Web Ontology Language (OWL) is a standard ontology language for the semantic web. “OWL includes conjunction, disjunction, existentially, and universally quantified variables, which can be used to carry out logical inferences and derive knowledge” (Kalibatiene and Vasilecas 2011). “Version 2 for OWL was adopted in October 2009 by W3C with minor revisions in December 2012” (Kendall and McGuinness 2019). OWL 2 introduced three sublanguages (called Profiles): OWL-EL, OWL-QL, OWL-RL. The OWL 2 profiles are trimmed down versions of OWL 2, since they trade expressive power for the efficiency of reasoning.

The **OWL 2 EL** captures the expressive power of ontologies with large number of properties and/or classes. OWL 2 EL performs reasoning in polynomial time with respect to the size of the ontology (Motik et al. 2009).

In applications with large instance data, query answering is the most important reasoning task. For such use cases, **OWL 2 QL** is used because QL implements conjunctive query answering using conventional relational database systems. In OWL 2 QL, reasoning can be performed in LOGSPACE with respect to the size of the assertions (Motik et al. 2009).

OWL 2 RL systems can be implemented using rule-based reasoning engines and are aimed at applications where scalable reasoning is required without sacrificing too much expressive power. The ontology consistency, class expression satisfiability, class expression subsumption, instance checking, and conjunctive query answering problems can be solved in time that is polynomial with respect to the size of the ontology (Motik et al. 2009).

SKOS: The Simple Knowledge Organization System (SKOS) data model is a W3C recommendation for sharing and linking knowledge organization systems via the web. SKOS is particularly useful for encoding knowledge organization systems like thesauri, classification schemes, subject

heading systems, and taxonomies (Isaac and Summers 2009). The SKOS data model, which is formally defined as an OWL ontology, represents a knowledge organization system as a concept scheme, consisting of a set of concepts. SKOS data are expressed as RDF triples and can be encoded using any RDF syntax (Isaac and Summers 2009).

KIF: The Knowledge Interchange Format is a W3C recommendation and computer-oriented language for the interchange of knowledge among disparate programs. KIF is logically comprehensive and has declarative semantics (i.e., the meaning of expressions in the representation can be understood without appeal to an interpreter for manipulating those expressions). It also provides for representation of meta-knowledge and non-monotonic reasoning rules. Lastly, it also provides for definitions of objects, functions, and relations.

DAML + OIL: DARPA Agent Markup Language + Ontology Inference Layer is a semantic markup language for Web resources. DAML + OIL provides a rich set of constructs with which to create machine readable and understandable ontologies. It is also a precursor to OWL and is rarely used in contemporary ontology engineering.

Ontology Engineering and Big Data

Ontology engineering faces the same scalability problems faced by most symbolic artificial intelligence systems. Ontologies are mostly built by humans (specifically a team of ontology and domain experts who work together). The increase in the number of instances in the ontology or broadening of scope of the ontology, i.e., adding parts of the domain previously unaddressed in the ontology (which result in additions and changes in the classes and properties of the ontology) are difficult to implement for ontologies that are continuously developed and improved. Ontologies have the capability to add new instances to the knowledge base without concern for the volume of data, as long as the correct concepts are

identified. The same cannot be said for restrictions or axioms in an ontology, with the possible exception of RDF, where descriptions can be translated into first order predicate calculus logical axioms. In practice though, large volumes of data need to be processed from their original form (either unstructured text or XML, HTML documents, etc.) into instances in the knowledge base. With the amount of data increasing rapidly, processing data from multiple sources to populate the instances in a knowledge base requires the application of automated methods in order to aid the development an ontology, i.e., ontology engineering.

Ontology Learning

The scaling up of an ontology to include large volumes of data (processed into instances) remains an ongoing problem, which is being fervently researched. Researchers often seek to automate either parts or all of the ontology engineering process in order eliminate or at least reduce the increasing workload. This process of automation is called ontology learning. Cimiano et al. (2009) and Asim et al. (2018) provide a comprehensive introduction to ontology learning; a generic architecture of such systems; and discusses current problems and challenges in the field. The general trend in ontology learning is focused on the acquisition of either taxonomic hierarchies or the construction of knowledge bases for existing ontologies. Such approaches do help resolve some of the scalability problems in ontology engineering.

OWL allows for modeling expressive axiomatizations. The field of automated axiom generation becomes a very important branch of ontology learning, because reasoning using axioms and restrictions is critical in inferring knowledge from an ontology, and as additional classes or instances are added to the ontology, there is a need to add new axioms that address the newly added data. Automated axiom generation, however, is a nascent and underexplored

area of research. There are some systems that can automatically generate disjointness axioms or use inductive logic programming to generate general axioms from schematic axioms (Cimiano et al. 2009; Asim et al. 2018). There are also methods that can extract rules from text documents (Dragoni et al. 2016). However, these methods require the rule to be explicitly stated (in natural language) in the text document. Truly automated axiom generation methods, that can learn to generate axioms based on the data in the knowledge base, still remain an unsolved and fervently researched problem.

Further Reading

- Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. *Database (Oxford)*, 2018, bay101. <https://doi.org/10.1093/database/bay101>.
- Cimiano, P., Mädche, A., Staab, S., & Völker, J. (2009). Ontology learning. In *Handbook on ontologies* (pp. 245–267). Berlin/Heidelberg: Springer.
- Dragoni, M., Villata, S., Rizzi, W., & Governatori, G. (2016, December). Combining NLP approaches for rule extraction from legal documents. In 1st workshop on mining and reasoning with legal texts (MIREL 2016).
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5–6), 907–928.
- Isaac, A., & Summers, E. (2009). Skos simple knowledge organization system. Primer, World Wide Web Consortium (W3C), 7.
- Kalibatiene, D., & Vasilecas, O. (2011). Survey on ontology languages. In *Perspectives in business informatics research* (pp. 124–141). Cham: Springer.
- Kendall, E., & McGuinness, D. (2019). *Ontology engineering (Synthesis lectures on the semantic web: theory and technology)*. Ying Ding and Paul Groth: Morgan & Claypool Editors.
- Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., & Lutz, C. (2009). OWL 2 web ontology language profiles. W3C recommendation, 27, 61.
- Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. <http://ftp.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf>.
- Smith, B., & Welty, C. (2001, October). Ontology: Towards a new synthesis. In *Formal ontology in information systems* (Vol. 10(3), pp. 3–9). Ogunquit: ACM Press.

Open Data

Alberto Luis García

Departamento de Ciencias de la Comunicación Aplicada, Facultad de Ciencias de la Información, Universidad Complutense de Madrid, Madrid, Spain

The link between Open Data and Big Data is articulated in relation to the need of public administrations to manage and analyze large volumes of data. The management of these data is based on the same technology used for the Big Data, but the difference with the Open Data lies in the origin of these data.

Today's democratic societies demand transparency in the management of their resources, so they demand open governments. At the same time, the European Union points to Open Data as a tool for innovation and growth.

Therefore, they have legislated the use of Open Data through portals where data repositories are hosted, following the workflow established with the use of Big Data: classification and quality control of information, processing architectures, and usability in the presentation and analysis of data.

Open Data is related as documents held by the public sector, by individuals or legal entities, for commercial or noncommercial purposes, provided that this use does not constitute a public administrative activity. In any case, for all this information is considered as Open Data must publish information in standard, open, and interoperable formats, allowing easier access and reuse. Open Data must complete the next sequence to get the major proposals and to convert the information in a public service to reuse them with strategical purposes:

1. Complete: It is necessary and reasonable attempt to cover the entire spectrum of data available on the subject that the data comes from. This step is crucial because it involves their success over the rest of the process. They must also be structured so that they can expand permanently and constantly updated so that

they can achieve the overall purpose of open access.

2. **Accessible:** This characteristic is the key in the relevance of Open Data. Accessibility should be universal and should be treated in a format, both search and output data, enabling all citizens, individually or commercial, can access and use them directly as a source of information.
3. **Free:** The system of free access and use should be provided for in the legislation itself to regulate public access to Open Data. However, the ultimate goal is to reuse data for commercial, social, economic, political, etc., to help integrate them into strategies and decision making in the executive staff of companies in the private and public sector. This feature underlies the legal possibility of allowing access to data that otherwise would violate the privacy rights.
4. **Nondiscriminatory:** In this respect, non-discrimination should occur in two ways. First, the possibility of universal access through systems and web pages that meet accessibility standards for anyone; on the other hand, the data must respect the particularities of gender, age, and religion and must meet the information needs for all social, religious, and ethnic groups that request.
5. **Nonproprietary:** data must be public and must not belong to any organization or private institution. The management must be controlled by government agencies in response to the regulations of each country in this regard. In any case, the defense of individual rights and personal freedoms of individuals and institutions should be honored and respected. This use of documents held by the public sector can be performed by individuals or legal entities, commercial or noncommercial purposes, provided such use does not constitute a public administrative activity. The ownership of the same, therefore, is the public sector, although the use thereof is regulated to establish and develop economic activities to private gain. In any case, the data reuse activities cannot be assumed to override the decision making in the public and make them look passed on all the citizens. In Spain, for example, use is regulated by Law 37/2007, of 16 November, on the reuse

of public sector information, in which it is assumed the philosophy of Directive 2003/98/EC, which states that “the use of such documents for other reasons set out either commercial or non-commercial purposes is reuse.”

In this sense, the Open Data principal agents (Users, Facilitators and Infomediaries) must be attentive to all processes respect these basic principles.

The Users (citizens or professionals) who demand information or new services initiate the process of action to Open Data access being the ultimate goal on which to structure the organization and management of data access. Data suppliers: only public administrations. Other kinds of suppliers must be integrated in the public regulation organism to ensure the correct use of them.

The Facilitators promote legal schemes and technical mechanisms making reuse possible.

The Infomediaries are the creators of the products and services based on the sources (students, professionals or public, private, or third-sector entities). The ultimate goal of Open Data is extracted from the same added value through the generation of applications and services that help solve specific demands from users; without this principle, there would be no value action on Open Data. Also, how to monetize this information service is an emerging economic value in businesses that need access to the peculiarities and needs of each individual to capitalize more effectively promoting every product and brand message.

Thus, the Open Data advantages are acting in this direction and are distributed in two main parts in response to social criteria: public and private benefits.

As far as public benefits are concerned, we could say that the main one relates to Open Government and transparency. There is a new trend based on transparency, participation, and collaboration, advocating a model of participatory democracy. The Open Data provides the necessary and sufficient information to create broader democratic models based on transparency in the access and visibility of data synergy, citizen participation in decision-making from the Open Data.

In the same direction we find public participation and integration. In this line of work, the use of

Open Data enables the improvement in the quality of monitoring of public policies, thus allowing greater participation and collaboration of citizens. The cooperation of citizens in the governance of cities from Open Data runs through two fundamental ways: access to public places of information (mainly web pages and apps) and by creating civilian institutions organizing the data management response to specific areas of interest. In this sense, citizens are called for greater participation in the governance of their cities and this means better management of Open Data in order not to limit the fundamental rights of access to public information.

Another property is the data quality and interoperability. The characteristics of the data must meet the following stipulations in order to achieve optimization and the basic characteristics of the information to use in a systematic and accessible for any user. The data must be **complete**: there should be no limitation and prior checking in introducing them, unless those issues that limit the legislation itself in defense of the rights of personal privacy; **primary**, that is, the data must be unprocessed and filtered as this is the core functionality that should contrast the infomediary forward to the specific objectives proposed. Also data must be **accessible**, avoiding any kind of restriction but should be a compulsory registration of users to have control over access to data. A line that contains the data that are available on the Web (in any format and according to the criteria of accessibility of the Web) must clearly exist. Another characteristic is that it must be **provided on time**, or what is the same, there must be a continuous input that allows the constant updating of the same. Success in Open Data is the principle of immediate upgrade in order to monetize the information immediately to obtain success always results in real time. The Open Data must be **processable** or structured so that it can work with them without special tools required data. Access to the data should be universal and not limited to specifications that prevent the normal processing of the same. For example, it would be mandatory to have data ready to work with them in any spreadsheet, place the image in a table that prevents default operability with such data. The last two

features are that Open Data must be **non-discriminatory**, limited by technical constraints or need for expensive high quality connections or technical access, so it is also convenient to introduce URLs identifying data from other websites and try to connect with data originating from other websites; and **no owners and license-free formats** as being public information should not be delivered in formats that benefit few software companies over others. For example, you can raise deliver data in CSV format instead of Excel.

Economic and employment growth: There is a clear tendency to use the Open Data as a clear way to develop effective new business models, which directly affects two fundamental aspects: a trend of increased employment of skilled labor and reducing costs by consolidating assets that affect the structural organization of companies delegating work through data with no charge. Contrasting examples of these benefits can be found in the following: subject MEPSIR [2006](#) report prepared by the European Commission, is producing a profit of up to € 47,000 million per year working with Open Data; in Spain, in particular, according to the Characterization Study infomediary Sector within the Provides Project and dated June 2012, there has been a profit between 330 and 550€ million and between 3600 and 4400 direct jobs. But since 2016, sector staffs have grown significantly: the total number of employees is estimated to be 14.3% higher (14,000–16,000 employees, in the most positive estimate). The collaboration between the public and private sectors in generating companies, services, and applications from the Open Data are leading to a change in the production model which is very evident. Part of the internal current expenditure of the public sector provides liquidity for public access technology infrastructure in the form of products, services, and therefore employment. The result that is occurring is a multiplier effect on the collaboration between the public and private sectors due to a recurring public contribution that gives value to public investment and generating profits in companies that request. In this sense, and within the European Union, it is creating a harmonization and standardization of Open Data Metadata for Open Apps proposals

from IEEE and JoinUP carried out in the Digital Agenda of the EU.

Social benefits, as transparency and civil government: In this sense, there are a historical background and legislative framework that in Europe, for example, began in 2003, with the Directive 2003/98/EC of the European Parliament and the Council, of 17 November 2003, on the reuse of public sector information. Each country has developed their own regulation in order to define the Technical Interoperability Standard for the Re-use of information resources.

The forms of reuse of data have changed from consultation services to a personalized manner through public agencies willing to do RSS content services that act as continuous indicators through scheduled alerts and filtered according to specific criteria. There is also the possibility of going through Web service that can presently offer regulated manners to all public bodies. In this case, being the main form of reuse through raw data.

The basic scheme for accessing documents reusable Open Data involves: (a) General basic mode (open-open access data); (b) License type in two ways: licenses-type “free” call is free information, and licenses-type specific, in which conditions are established for reuse information. In any case, inaccessible involved data access limitation should never exist. (c) Re-request: a general method for the request of documents. This way is reserved for application data and subject to more specific characteristics which may be confronted with some aspect of the general regulatory standards conditions.

In any case, you should take in all the International standards inquiry Industrial Property, including the ST 36, ST 66, and ST 86. In addition, we must take into account the Data Protection Act and have control over the use of the legal basis for the re-obligations that should aim to improve the quality of data at all times and consultation. To do this, regulators should give accessibility and data entry in the catalogue to both the user and the infomediary.

The general conditions applicable to all re-users go through basic concepts, but essential as not to distort the data, cite the source, and update the date of consultation at all times to comply with the reliability thereof.

Moreover, all agents must retain re-user meta-data to certify at all times, the source thereof. All these principles need to be articulated and regulated as an emergency that could produce a conflict in the use of Open Data. The legislation must overcome geographical barriers since the use thereof affects globally, so that international organizations are getting involved in legislation.

Definitively, Open Data influences the way that people relate to institutions in the future. However, there are currently a number of barriers that need to be solving for the sake of process consistency. Among the main barriers we are currently facing, there is a lack of commitment from the Public Sector to promote the reuse of Open Data, which could be improved with proper training of public employees. In that direction, there are difficulties held by public entities to carry out Open Data strategies and it is necessary to disseminate benefits of Open Data. Also, it is necessary for data opening up: identification, cataloguing, and classification, so it is very important to promote citizens’ and private organizations’ participation in the reuse of data.

Further Reading

- Estudio de Caracterización del Sector Infomediario en España. <http://www.ontsi.red.es/ontsi/es/estudios-informes/estudio-de-caracterizaci%C3%B3n-del-sector-infomediario-en-esp%C3%B1a-edici%C3%B3n-2012>. Accessed Aug 2014.
- Estudio de Caracterización del Sector Infomediario en España. http://datos.gob.es/sites/default/files/Info_sector%20infomediario_2012_vfr.pdf. Accessed Aug 2014.
- Estudio de Caracterización del Sector Infomediario en España. <https://www.ontsi.red.es/sites/ontsi/files/2020-06/PresentationCharacterizationInfomediarySector2020.pdf>. Accessed Aug 2020.
- EU Data Protection Directive 95/46/EC. http://europa.eu/legislation_summaries/information_society/data_protection/114012_es.htm. Accessed Aug 2014.
- Industrial Property, including the ST 36, ST 66 and ST 86. <http://www.wipo.int/export/sites/www/standards/en/pdf/03-96-01.pdf>. Accessed Aug 2014.
- Law 37/2007, of 16 November, on the reuse of public sector information. http://www.boe.es/diario_boe/txt.php?id=BOE-A-2007-19814. Accessed Aug 2014.
- MEPSIR Report. (2006). <http://www.cotec.es/index.php/pagina/publications/new-additions/show/id/952/titulo/reutilizacion-de-la-informacion-del-sector-publico%2D%2D2011>. Accessed Aug 2014.

Open-Source Software

Marc-David L. Seidel

Sauder School of Business, University of British Columbia, Vancouver, BC, Canada

Open-source software refers to computer software where the copyright holder provides anybody the right to edit, modify, and distribute the software free of charge. The initial creation of such software spawned the open-source movement. Frequently the only limitation on the intellectual property rights are that any subsequent changes made by others are required to be made with similarly open intellectual property rights. Such software is often developed in an open collaborative manner by a Community Form (C-form) organization. A large percentage of the internet infrastructure is operated utilizing such software which handles the majority of networking, web serving, e-mail, and network diagnostics. With the spread of the internet, the volume of user generated data has expanded exponentially, and open-source software to manage and analyze big data has flourished through open-source big data projects. This entry explains the history of open-source software, the typical organizational structure used to create such software, prominent project examples of the software focused on managing and analyzing big data, and the future evolution suggested by current research on the topic.

History of Open-Source Software

Two early software projects leading to the modern-day open-source software growth were at the Massachusetts Institute of Technology (MIT) and the University of California at Berkeley. The Free Software Foundation, created by Richard Stallman of the MIT Artificial Intelligence Lab, was launched as a nonprofit organization to promote the development of free software. Stallman is credited with creating the term “copyleft” and created the GNU operating system as an operating

system composed entirely of free software. The free BSD Unix operating system was developed by Bill Jolitz of the University of California at Berkeley Computer Science Research Group and served as the basis for many later Unix operating system releases. Many open-source software projects were unknown outside of the highly technical computer science community. Stallman’s GNU was later popularized by Linus Torvalds, a Finish computer science student, who released a Linux kernel based upon the earlier work. The release of Linux triggered substantial media attention for the open-source movement when an internal Microsoft strategy document, dubbed the Halloween Documents, was leaked. It outlined Microsoft’s perception of the threat of Linux to Microsoft’s dominance of the operating system market. Linux was portrayed in the mass media as a free alternative to the Microsoft Windows operating system. Eric S. Raymond and Bruce Perens further formalized open source as a development method by creating the Open Source Initiative in 1998. By 1998, open-source software routed 80% of the e-mail on the internet. It has continued to flourish to the modern day being responsible for a large number of software and information-based products today produced by the open-source movement.

C-form Organizational Architecture

The C-form organizational architecture is the primary organizational structure for open-source development projects. A typical C-form has four common organizing principles. First, there are informal peripheral boundaries for developers. Contributors can participate as much or as little as they like and join or leave a project on their own. Second, many contributors receive no financial compensation at all for their work, yet some may have employment relationships with more traditional organizations which encourage their participation in the C-form as part of their regular job duties. Third, C-forms focus on information-based product, of which software is a major subset. Since the product of a typical C-form is information based, it can be replicated with minimal

effort and cost. Fourth, typical C-forms operate with a norm of open transparent communication. The primary intellectual property of an open-source C-form is the software code. This, by definition, is made available for any and all to see, use, and edit.

Prominent Examples of Open-Source Big Data Projects

Apache Cassandra is a distributed database management system originally developed by Avinash Lakshman and Prashant Malik at Facebook as a solution to handle searching an inbox. It is now developed by the Apache Software Foundation, a distributed community of developers. It is designed to handle large amounts of data distributed across multiple datacenters. It has been recognized by University of Toronto researchers as having leading scalability capabilities.

Apache CouchDB is a web-focused database system originally developed by Damien Katz, a former IBM developer. Similar to Apache Cassandra, it is now developed by the Apache Software Foundation. It is designed to deal with large amounts of data through multi-master replication across multiple locations.

Apache Hadoop is designed to store and process large-scale datasets using multiple clusters of standardized low-level hardware. This technique allows for parallel processing similar to a super-computer but using mass market off the shelf commodity computing systems. It was originally developed by Doug Cutting and Mike Cafarella. Cutting was employed at Yahoo, and Cafarella was a Masters student at the University of Washington at the time. It is now developed by the Apache Software Foundation. It serves a similar purpose as Storm.

Apache HCatalog is a table and storage management layer for Apache Hadoop. It is focused on assisting grid administrators with managing large volumes of data without knowing exactly where the data is stored. It provides relational views of the data, regardless of what the source storage location is. It is developed by the Apache Software Foundation.

Apache Lucene is an information retrieval software library which tightly integrates with search engine projects such as Elasticsearch. It provides full text indexing and searching capabilities. It treats all document formats similarly by extracting textual components and as such is independent of file format. It is developed by the Apache Software Foundation and released under the Apache Software License.

D3.js is a data visualization package originally created by Mike Bostock, Jeff Heer, and Vadim Ogievetsky who worked together at Stanford University. It is now licensed under the Berkeley Software Distribution (BSD) open-source license. It is designed to graphically represent large amounts of data and is frequently used to generate rich graphs and for map making.

Drill is a framework to support distributed applications for data intensive analysis of large-scale datasets in a self-serve manner. It is inspired by Google's BigQuery infrastructure service. The stated goal for the project is to scale to 10,000 or more servers to make low-latency queries of petabytes of data in seconds in a self-service manner. It is being incubated by Apache currently. It is similar to Impala.

ElasticSearch is a search server that provides near real-time full-text search engine capabilities for large volumes of documents using a distributed infrastructure. It is based upon Apache Lucene and is released under the Apache Software License. It spawned a venture-funded company in 2012 created by the people responsible for Elasticsearch and Apache Lucene to provide support and professional services around the software.

Impala is an SQL query engine which enables massively parallel processing of search queries on Apache Hadoop. It was announced in 2012 and moved out of beta testing in 2013 to public availability. It is targeted at data analysts and scientists who need to conduct analysis on large-scale data without reformatting and transferring the data to a specialized system or proprietary format. It is released under the Apache Software License and has professional support available from the venture-funded Cloudera. It is similar to Drill.

Julia is a technical computing high-performance dynamic programming language with a focus on distributed parallel execution with high numerical accuracy using an extensive mathematical function library. It is designed to use a simple syntax familiar to many developers of older programming languages while being updated to be more effective with big data. The aim is to speed development time by simplifying coding for parallel processing support. It was first released in 2012 under the MIT open-source license after being originally developed starting in 2009 by Alan Edelman (MIT), Jeff Bezanson (MIT), Stefan Karpinski (UCSB), and Viral Shah (UCSB).

Kafka is a distributed, partitioned, replicated message broker targeted on commit logs. It can be used for messaging, website activity tracking, operational data monitoring, and stream processing. It was originally developed by LinkedIn and released open source in 2011. It was subsequently incubated by the Apache Incubator and as of 2012 is developed by the Apache Software Foundation.

Lumify is a big data analysis and visualization platform originally targeted to investigative work in the national security space. It provides real-time graphical visualizations of large volumes of data and automatically searches for connections between entities. It was originally created by Altamira Technologies Corporation and then released under the Apache License in 2014.

MongoDB is a NoSQL document focused database focused on handling large volumes of data. The software was first developed in 2007 by 10gen. In 2009, the company made the software open source and focused on providing professional services for the integration and use of the software. It utilizes a distributed file storage, load balancing, and replication system to allow quick ad hoc queries of large volumes of data. It is released under the GNU Affero General Public License and uses drivers released under the Apache License.

R is a technical computing high-performance programming language focused on statistical analysis and graphical representations of large

datasets. It is an implementation of the S programming language created by Bell Labs' John Chambers. It was created by Ross Ihaka and Robert Gentleman at the University of Auckland. It is designed to allow multiple processors to work on large datasets. It is released under the GNU License.

Scribe is a log server designed to aggregate large volumes of server data streamed in real time from a high volume of servers. It is commonly described as a scaling tool. It was originally developed by Facebook and then released in 2008 using the open-source Apache License.

Spark is a data analytic cluster computing framework designed to integrate with Apache Hadoop. It has the capability to cache large datasets in memory to interactively analyze the data and then extract a working analysis set to further analyze quickly. It was originally developed at the University of California at Berkeley AMPLab and released under the BSD License. Later it was incubated in 2013 at the Apache Incubator and released under the Apache License. Major contributors to the project include Yahoo and Intel.

Storm is a programming library focused on real-time storage and retrieval of dynamic object information. It allows complex querying across multiple database tables. It handles unbound streams of data in an instantaneous manner allowing real-time analytics of big data and continuous computation. The software was originally developed by Canonical Ltd., also known for the Ubuntu Linux operating system, and is released under the GNU Lesser General Public License. It is similar to Apache Hadoop but with a more real-time and less batch-focused nature.

The Future

The majority of open-source software focused on big data applications has primarily been targeting web-based big data sources and corporate data analytics. Current developments suggest a shift toward more analysis of real-world data as sensors spread more widely into everyday use by mass

market consumers. As consumers provide more and more data passively through pervasive sensors, the open-source software used to manage and understand big data appears to be shifting toward analyzing a wider variety of big data sources. It appears likely that the near future will provide more open-source software tools to analyze real-world big data such as physical movements, biological data, consumer behavior, health metrics, and voice content.

Cross-References

- [Crowdsourcing](#)
- [Google Flu](#)
- [Wikipedia](#)

Further Reading

- Bretthauer, D. (2002). Open source software: A history. *Information Technology and Libraries*, 21(1), 3–11.
- Lakhani, K. R., & von Hippel, E. (2003). How open source software works: ‘Free’ user-to-user assistance. *Research Policy*, 32(6), 923–943.
- Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498, 255–260.
- McHugh, J. (1998, August). For the love of hacking. *Forbes*.
- O’Mahony, S., & Ferraro, F. (2007). The emergence of governance on an open source project. *Academy of Management Journal*, 50(5), 1079–1106.
- Seidel, M.-D. L., & Stewart, K. (2011). An initial description of the C-form. *Research in the Sociology of Organizations*, 33, 37–72.
- Shah, S. K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science*, 52(7), 1000–1014.

Parallel Processing

► [Multiprocessing](#)

Participatory Health and Big Data

Muhiuddin Haider, Yessenia Gomez and Salma Sharaf
School of Public Health Institute for Applied Environmental Health, University of Maryland, College Park, MD, USA

The personal data landscaped has changed drastically with the rise of social networking sites and the Internet. The Internet and social media sites have allowed for the collection of large amounts of personal data. Every keystroke typed, website visited, Facebook post liked, Tweet posted, or video shared becomes part of a user's digital history. A large net is cast collecting all the personal data into big data sets that may be subsequently analyzed. This type of data has been analyzed for years by marketing firms through the use of algorithms that analyze and predict consumer purchasing behavior. The digital history of an individual paints a clear picture about their influence in the community and their

mental, emotional, and financial state, and much about an individual can be learned through the tracking of his or her data. When big data is fine-tuned, it can benefit the people and community at large. Big data can be used to track epidemics, and its analysis can be used in the support of patient education, treatment of at-risk individuals, and encouragement of participatory community health. However, with the rise of big data comes concern about the security of health information and privacy.

There are advantages and disadvantages to casting large data nets. Collecting data can help organizations learn about individuals and communities at large. Following online search trends and collecting big data can help researchers understand health problems currently facing the studied communities and can similarly be used to track epidemics. For example, increases in Google searches for the term flu have been correlated with an increase in flu patient visits to emergency rooms. In addition, a 2008 Pew study revealed that 80% of Internet users use the Internet to search for health information. Today, many patients visit doctors after having already searched their symptoms online. Furthermore, more patients are now using the Internet to search health information, seek medical advice, and make important medical decisions. The rise of the Internet has led to more patient engagement and participation in health.

Technology has also encouraged participatory health through an increase in interconnectedness. Internet technology has allowed for constant access to medical specialists and support groups for people suffering from diseases or those searching for health information. The use of technology has allowed individuals to take control of their own health, through the use of online searches and the constant access to online health records and tailored medical information. In the United States, hospitals are connecting individuals to their doctors through the use of online applications that allow patients to email their doctors, check prescriptions, and look at visit summaries from anywhere where they have an Internet connection. The increase in patient engagement has been seen to play a major role in promotion of health and improvement in quality of healthcare.

Technology has also helped those at risk of disease seek treatment early or be followed carefully before contracting a disease. Collection of big data has helped providers see health trends in their communities, and technology has allowed them to reach more people with targeted health information. A United Nations International Children's Emergency Fund (UNICEF) project in Uganda asked community members to sign up for U-report, a text-based system that allows individuals to participate in health discussions through weekly polls. This system was implemented to connect and increase communication between the community and the government and health officials. The success of the program helped UNICEF prevent disease outbreaks in the communities and encouraged healthy behaviors. U-report is now used in other countries to help mobilize communities to play active roles in their personal health.

Advances in technology have also created wearable technology that is revolutionizing participatory health. Wearable technology is a category of devices that are worn by individuals and are used to track data about the individuals, such as health information. Examples of wearable technology are wrist bands that collect information about the individual's global positioning system (gps) location, amount of daily exercise, sleep

patterns, and heart rate. Wearable technology enables users to track their health information, and some wearable technology even allows the individual to save their health information and share it with their medical providers. Wearable technology encourages participatory health, and the constant tracking of health information and sharing with medical providers allow for more accurate health data collection and tailored care. The increase in health technology and collection and analysis of big data has led to an increase in participatory health, better communication between individuals and healthcare providers, and more tailored care.

Big data collected from these various sources, whether Internet searches, social media sites, or participatory health through applications and technology, strongly influences our modern health system. The analysis of big data has helped medical providers and researchers understand health problems facing their communities and develop tailored programs to address health concerns, prevent disease, and increase community participatory health. Through the use of big data technology, providers are now able to study health trends in their communities and communicate with their patients without scheduling any medical visits. However, big data also creates concern for the security of health information.

There are several disadvantages to the collection of big data. One being that not all the data collected is significant and much of the information collected may be meaningless. Additionally, computers lack the ability to interpret information the way humans do, so something that may have multiple interpretations may be misinterpreted by a computer. Therefore, data may be flawed if simply interpreted based on algorithms, and any decisions regarding the health of the communities that were made based on this inaccurate data would also be flawed. Of greater concern is the issue of privacy with regards to big data. Much of the data is collected automatically based on people's online searches and Internet activities, so the question arises as to whether people have the right to choose what data is collected about them. Questions that arise

regarding big data and health include how long is personal health data saved? Will data collected be used against individuals? How will the Health Insurance Portability and Accountability Act (HIPPA) change with the incorporation of big data in medicine? Will data collected determine insurance premiums? Privacy concerns need to be addressed before big health data, health applications, and wearable technology become a security issue.

Today, big data can help health providers better understand their target populations and can lead to an increase in participatory health. However, concerns arise about the safety of health information that is automatically collected in big data sets. With this in mind, targeted data collected may be a more beneficial method for data collection with regard to health. All these concerns need to be addressed today as the use of big data in health becomes more commonplace.

Cross-References

- ▶ Epidemiology
- ▶ Online Advertising
- ▶ Patient-Centered (Personalized) Health
- ▶ PatientsLikeMe
- ▶ Prevention

Further Reading

- Eysenbach, G. (2008). Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness. *Journal of Medical Internet Research*, 10(3), e22. <https://doi.org/10.2196/jmir.1030>.
- Gallant, L. M., Irizarry, C., Boone, G., & Kreps, G. (2011). Promoting participatory medicine with social media: New media applications on hospital websites that enhance health education and e-patients' voices. *Journal of Participatory Medicine*, 3, e49.
- Gallivan, J., Kovacs Burns, K. A., Bellows, M., & Eigenseher, C. (2012). The many faces of patient engagement. *Journal of Participatory Medicine*, 4, e32.
- Lohr, S. (2012). The age of big data. *The New York Times*. Revolutionizing social mobilization, monitoring and response efforts. (2012) UNICEF [video file]. Retrieved from <https://www.youtube.com/watch?v=gRczMq1Dn10>.
- The promise of personalized medicine. (2007, Winter). *NIH Medline Plus*, pp. 2–3.

Patient Records

Barbara Cook Overton

Communication Studies, Louisiana State University, Baton Rouge, LA, USA

Communication Studies, Southeastern Louisiana University, Hammond, LA, USA

Patient records have existed since the first hospitals were opened. Early handwritten accounts of patients' hospitalizations were recorded for educational purposes but most records were simply tallies of admissions and discharges used to justify expenditures. Standardized forms would eventually change how patient care was documented. Content shifted from narrative to numerical descriptions, largely in the form of test results. Records became unwieldy as professional guidelines and malpractice concerns required more and more data be recorded. Patient records are owned and maintained by individual providers, meaning multiple records exist for most patients. Nonetheless, the patient record is a document meant to ensure continuity of care and is a communication tool for all providers engaged in a patient's current and future care. Electronic health records may facilitate information sharing, but that goal is largely unrealized.

Modern patient records evolved with two primary goals: facilitating fiscal justification and improving medical education. Early hospitals established basic rules to track patient admissions, diagnoses, and outcomes. The purpose was largely bureaucratic: administrators used patient tallies to justify expenditures. As far back as 1737, Berlin surgeons were required to note patients' conditions each morning and prescribe lunches accordingly (e.g., soup was prescribed for patients too weak to chew). The purpose, according to Volker Hess and Sophie Ledebur, was helping administrators track the hospital's food costs and had little bearing on actual patient care. In 1791, according to Eugenia Siegler in her analysis of early medical recordkeeping, the New York Board of Governors required complete patient logs along with lists of prescribed medications, but no

descriptions of the patients' conditions. Formally documenting the care that individual patients received was fairly uncommon in American hospitals at that time. It was not until the end of the nineteenth century that American physicians began recording the specifics of daily patient care for all patients. Documentation in European hospitals, by contrast, was much more complete. From the mid-eighteenth century on, standardized medical forms were widely used to record patients' demographic data, their symptoms, treatments, daily events, and outcomes. By 1820, these forms were collected in preprinted folders with multiple graphs and tables (by contrast, American hospitals would not begin using such forms until the mid-1860s). Each day, physicians in training were tasked with transcribing medical data into meaningful narratives, describing patterns of disease progression. The resulting texts became valuable learning tools. Similar narratives were compiled by American physicians and used for medical training as well. In 1805, Dr. David Hosack had suggested recording the specifics of particularly interesting cases, especially those holding the greatest educational value for medical students. The New York Board of Governors agreed and mandated compiling summary reports in casebooks. As Siegler noted, there were very few reports written at first: the first casebook spanned 1810–1834. Later, as physicians in training were required to write case reports in order to be admitted to their respective specialties, the number of documented cases grew. Eventually, reports were required for all patients. The reports, however, were usually written retrospectively and in widely varying narrative styles.

Widespread use of templates in American hospitals helped standardize patient records, but the resulting quantitative data superseded narrative content. By the start of the twentieth century, forms guaranteed documentation of specific tasks like physical exams, histories, orders, and test results. Graphs and tables dominated patient records and physicians' narrative summaries began disappearing. The freestyle narrative form that had previously comprised the bulk of the patient record allowed physicians to write as much or as little as they wished. Templates left

little room for lengthy narratives, no more than a few inches, so summary reports gave way to brief descriptions of pertinent findings. As medical technology advanced, according to Siegler, the medical record became more complicated and cumbersome with the addition of yet more forms for reporting each new type of test (e.g., chemistry, hematology, and pathology tests). While most physicians kept working notes on active patients, these scraps of paper notating observations, daily tasks, and physicians' thoughts seldom made their way into the official patient record. The official record emphasized tests and numbers, as Siegler noted, and this changed medical discourse: interactions and care became more data driven. Care became less about the totality of the patient's experience and the physician's perception of it. Nonetheless, patient records had become a mainstay and they did help ensure continuity of care. Despite early efforts at a unifying style, however, the content of patient records still varied considerably.

Although standardized forms ensured certain events would be documented, there were no methods to ensure consistency across documentations or between providers. Dr. Larry Weed proposed a framework in 1964 to help standardize recording medical care: SOAP notes. SOAP notes are organized around four key areas: subjective (what patients say), objective (what providers observe, including vital signs and lab results), assessment (diagnosis), and plan (prescribed treatments). Other standardized approaches have been developed since then. The most common charting formats today, in addition to SOAP notes, include narrative charting, APIE charting, focus charting, and charting by exception. Narrative charting, much as in the early days of patient record-keeping, involves written accounts of patients' conditions, treatments, and responses and is documented in chronological order. Charts include progress notes and flow sheets which are multi-column forms for recording dates, times, and observations that are updated every few hours for inpatients and upon each subsequent outpatient visit. They provide an easy-to-read record of change over time; however their limited space cannot take the place of more complete

assessments, which should appear elsewhere in the patient record. APIE charting, similar to SOAP notes, involves clustering patient notes around assessment (both subjective and objective findings), planning, implementation, and evaluation. Focus charting is a more concise method of inpatient recording and is organized by keywords listed in columns. Providers note their actions and patients' responses under each keyword heading. Charting by exception involves documenting only significant changes or events using specially formatted flow sheets. Computerized charting, or electronic health records (EHR), combines several of the above approaches but proprietary systems vary widely. Most hospitals and private practices are migrating to EHRs, but the transition has been expensive, difficult, and slower than expected. The biggest challenges include interoperability issues impeding data sharing, difficult-to-use EHRs, and perceptions that EHRs interfere with provider-patient relationships.

Today, irrespective of the charting format used, patient records are maintained according to strict guidelines. Several agencies publish recommended guidelines including the American Association of Nurses, the American Medical Association (AMA), the Joint Commission of Accreditation of Healthcare Organizations (JCAHO), and the Centers for Medicare and Medicaid Services (CMS). Each regards the medical record as a communication tool for everyone involved in the patient's current and future care. The primary purpose of the medical record is to identify the patient, justify treatment, document the course of treatment and results, and facilitate continuity of care among providers. Data stored in patient records have other functions; aside from ensuring continuity of care, data can be extracted for evaluating the quality of care administered, released to third-party payers for reimbursement, and analyzed for clinical research and/or epidemiological studies. Each agency's charting guidelines require certain fixed elements in the patient record: the patient's name, address, birthdate, attending physician, diagnosis, next of kin, and insurance provider. The patient record also contains physicians' orders and progress notes, as well as medication lists, X-ray records, laboratory

tests, and surgical records. Several agencies require the patient's full name, birthdate, and a unique patient identification number appear on each page of the record, along with the name of the attending physician, date of visit or admission, and the treating facility's contact information. Every entry must be legibly signed or initialed and date/time stamped by the provider.

The medical record is a protected legal document and because it could be used in a malpractice case, charting takes on added significance. Incomplete, confusing, or sloppy patient records could signal poor medical care to a jury, even in the absence of medical incompetence. For that reason, many malpractice insurers require additional documentation above and beyond what professional agencies recommend. For example, providers are urged to: write legibly in permanent ink, avoid using abbreviations, write only objective/quantifiable observations and use quotation marks to set apart patients' statements, note communication between all members of the care team while documenting the corresponding dates and times, document informed consent and patient education, record every step of every procedure and medication administration, and chart instances of patients' noncompliance or lack of cooperation. Providers should avoid writing over, whiting out, or attempting to erase entries, even if made in error – mistakes should be crossed through with a single line, dated, and signed. Altering a patient chart after the fact is illegal in many states, so corrections should be made in a timely fashion and dated/signed. Leaving blank spaces on medical forms should be avoided as well; if space is not needed for documenting patient care, providers are instructed to draw a line through the space or write "N/A." The following should also be documented to ensure both good patient care and malpractice defense: the reason for each visit, chief complaint, symptoms, onset and duration of symptoms, medical and social history, family history, both positive and negative test results, justifications for diagnostic tests, current medications and doses, over-the-counter and/or recreational drug use, drug allergies, any discontinued medications and reactions, medication renewals or dosage changes, treatment

recommendations and suggested follow-up or specialty care, a list of other treating physicians, a “rule-out” list of considered but rejected diagnoses, final definitive diagnoses, and canceled or missed appointments.

Patient records contain more data than ever before because of professional guidelines, malpractice-avoidance strategies, and the ease of data entry many EHRs make possible. The result is that providers are experiencing data overload. Many have difficulty wading through mounds of data, in either paper or electronic form, to discern important information from insignificant attestations and results. While EHRs are supposed to make searching for data easier, many providers lack the needed skills and time to search for and review patients’ medical records. Researchers have found some physicians rely on their own memories or ask patients about previous visits instead of searching for the information themselves. Other researchers have found providers have trouble quickly processing the amount of quantitative data and graphs in most medical records. Donia Scott and colleagues, for example, found that providers given narrative summaries of patient records culled from both quantitative and qualitative data performed better on questions about patients’ conditions than those providers given complete medical records, and did so in half the time. Their findings highlight the importance of narrative summaries that should be included in patients’ records. There is a clear need for balancing numbers with words in ensuring optimal patient care.

Another important issue is ownership of and access to patient records. For each healthcare provider and/or medical facility involved in a patient’s care, there is a unique patient record owned by that provider. With patients’ permission, those records are frequently shared among providers. The Health Insurance Portability and Accountability Act (HIPAA) protects the confidentiality of patient data, but patients, guardians or conservators of minor or incompetent patients, and legal representatives of deceased patients may request access to records. Providers in some states can withhold records if, in the providers’ judgment, releasing information could be detrimental

to patients’ well-being or cause emotional or mental distress. In addition to HIPAA mandates, many states have strict confidentiality laws restricting the release of HIV test results, drug and alcohol abuse treatment, and inpatient mental health records. While HIPAA guarantees patient access to their medical records, providers can charge copying fees. Withholding records because a patient cannot afford to pay for them is prohibited in many states because it could disrupt the continuity of care. HIPAA also allows patients the right to amend their medical records if they believe mistakes have been made. While providers are encouraged to maintain records in perpetuity, there are not requirements that they do so. Given the costs associated with data storage, both on paper and electronically, many providers will only maintain charts on active patients. Many inactive patients, those who have not seen a given provider in 8 years, will likely have their records destroyed. Additionally, many retiring physicians typically only maintain records for 10 years. Better data management capabilities will inevitably change these practices in years to come.

While patient records have evolved to ensure continuity of patient care, many claim the current form that records have taken facilitates billing over communication concerns. Many EHRs, for instance, are modeled after accounting systems: providers’ checkbox choices of diagnoses and tests are typically categorized and notated in billing codes. Standardized forms are also designed with billing codes in mind. Diagnosis codes are reported in the International Statistical Classification of Diseases and Related Health Problems terminology, commonly referred to as ICD. The World Health Organization maintains this coding system for epidemiological, health management, and research purposes. Billable procedures and treatments administered in the United States are reported in Current Procedural Terminology (CPT) codes. The AMA owns this coding schema and users must pay a yearly licensing fee for the CPT codes and codebooks, which are updated annually. Critics claim this amounts to a monopoly, especially given HIPAA, CMS, and most insurance companies require CPT-coded data to

satisfy reporting requirements and for reimbursement. CPT-coded data may impact patients' ability to decipher and comprehend their medical records, but the AMA does have a limited search function on its website for non-commercial use allowing patients to look up certain codes.

Patient records are an important tool ensuring continuity of care, but data-heavy records are cumbersome and often lacking narrative summaries which have been shown to enhance providers' understanding of patients' histories and inform better medical decision-making. Strict guidelines and malpractice concerns produce thorough records that while ensuring complete documentation, sometimes impede providers' ability to discern important from less significant past findings. Better search and analytical tools are needed for managing patient records and data.

Cross-References

- ▶ [Electronic Health Records \(EHR\)](#)
- ▶ [Health Care Delivery](#)
- ▶ [Health Informatics](#)
- ▶ [Patient-Centered \(Personalized\) Health](#)

Further Reading

- American Medical Association. *CPT – current procedural terminology*. <http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page>. Accessed Oct 2014.
- Christensen, T., & Grimsno, A. (2008). Instant availability of patient records, but diminished availability of patient information: A multi-method study of GP's use of electronic health records. *BMC Medical Informatics and Decision Making*, 8(12), doi:10.1186/1472-6947-8-12
- Hess, V., & Ledebur, S. (2011). Taking and keeping: A note on the emergence and function of hospital patient records. *Journal of the Society of Archivists*, 32, 1.
- Lee, J. Interview with Lawrence Weed, MD – *The father of the problem-oriented medical record looks ahead*. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2911807/>. Accessed Oct 2014.
- Medical Insurance Exchange of California. *Medical record documentation for patient safety and physician defensibility*. <http://www.miec.com/Portals/0/pubs/MedicalRec.pdf>. Accessed Oct 2014.

Scott, D., et al. (2013). Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories. *Patient Education and Counseling*, 92, 153–159.

Siegler, E. (2010). The evolving medical record. *Annals of Internal Medicine*, 153, 671–677.

Patient-Centered (Personalized) Health

Barbara Cook Overton

Communication Studies, Louisiana State University, Baton Rouge, LA, USA

Communication Studies, Southeastern Louisiana University, Hammond, LA, USA

Patient-centered health privileges patient participation and results in tailored interventions incorporating patients' needs, values, and preferences. Although this model of care is preferred by patients and encouraged by policy makers, many healthcare providers persist in using a biomedical approach which prioritizes providers' expertise and downplays patients' involvement. Patient-centered care demands collaborative partnerships and quality communication, both requiring more time than is generally available during medical exams. While big data may not necessarily improve patient-provider communication, it can facilitate individualized care in several important ways.

The concept of patient-centered health, although defined in innumerable ways, has gained momentum in recent years. In 2001, the Institute of Medicine (IOM) issued a report recommending healthcare institutions and providers adopt six basic tenets: safety, effectiveness, timeliness, efficiency, equity, and patient-centeredness. Patient-centeredness, according to the IOM, entails delivering quality health care driven by patients' needs, values, and preferences. The Institute for Patient- and Family-Centered Care expands the IOM definition by including provisions for shared decision-making, planning, delivery, and evaluation of health care that is situated in partnerships comprising patients, their families, and providers.

The concept is further elucidated in terms of four main principles: respect, information sharing, participation, and collaboration. According to the Picker Institute, patient-centered care encompasses seven basic components: respect, coordination, information and education, physical comfort, emotional support, family involvement, and continuity of care. All of the definitions basically center on two essential elements: patient participation in the care process and individualized care.

The goal of patient-centered care, put forth by the IOM, is arguably a return to old-fashioned medicine. Dr. Abraham Flexner, instrumental in revamping physician training during the 1910s and 1920s, promoted medical interactions that were guided by both clinical reasoning and compassion. He encouraged a biopsychosocial approach to patient communication, which incorporates patients' feelings, thoughts, and expectations. Scientific and technological advances throughout the twentieth century, however, gradually shifted medical inquiry away from the whole person and towards an ever-narrowing focus on symptoms and diseases. Once the medical interview became constricted, scientific, and objective, collaborative care gave way to a provider-driven approach. The growth of medical specialties (like cardiology and gastroenterology) further compounded the problem by reducing patients to collections of interrelated systems (such as circulatory and digestive). This shift to specialty care coincided with fewer providers pursuing careers in primary care, the specialty most inclined to adopt a patient-centered perspective. The resulting biomedical model downplays patient participation while privileging provider control and expertise. Although a return to patient-centered care is being encouraged, many providers persist in using a biomedical approach. Some researchers fault patients for not actively co-constructing the medical encounter, while others blame medical training that de-emphasizes relationship development and communication skills.

Several studies posit quality communication as the single most important component necessary for delivering patient-centered care. Researchers find patient dissatisfaction is associated with

providers who are insensitive to or misinterpret patients' socio-emotional needs, fail to express empathy, do not give adequate feedback or information regarding diagnoses and treatment protocols, and disregard patients' input in decision-making. Patients who are dissatisfied with providers' communication are less likely to comply with treatment plans and typically suffer poorer outcomes. Conversely, patients satisfied with the quality of their providers' communication are more likely to take medications as prescribed and adhere to recommended treatments. Satisfied patients also have lower blood pressure and better overall health. Providers, however, routinely sacrifice satisfaction for efficiency, especially in managed care contexts.

Many medical interactions proceed according to a succinct pattern that does not prioritize patients' needs, values, and preferences. The asymmetrical nature of the provider-patient relationship preferences providers' goals and discourages patient participation. Although patients expect to have all or most of their concerns addressed, providers usually pressure them to focus on one complaint per visit. Providers also encourage patients to get to the point quickly, which means patients rarely speak without interruption or redirection. While some studies note patients are becoming more involved in their health care by offering opinions and asking questions, others find ever-decreasing rates of participation during medical encounters. Studies show physicians invite patients to ask questions in fewer than half of exams. Even when patients do have concerns, they rarely speak up because they report feeling inhibited by asymmetrical relationships: many patients simply do not feel empowered to express opinions, ask questions, or assert goals. Understandably, communication problems stem from these hierarchical differences and competing goals, thereby making patient-centered care difficult.

There are several other obstacles deterring patient-centered communication and care. While medical training prioritizes the development of clinical skills over communication skills, lack of time and insufficient financial reimbursement are the biggest impediments to patient-centered care.

The “one complaint per visit” approach to health care means most conversations are symptom specific, with little time left for discussing patients’ overall health goals. Visits should encompass much broader health issues, moving away from the problem presentation/treatment model while taking each patient’s unique goals into account. The goal of patient-centered care is further compromised by payment structures incentivizing quick patient turnaround over quality communication, which takes more time than is currently available in a typical medical encounter. Some studies, however, suggest that patient-centered communication strategies, like encouraging questions, co-constructing diagnoses, and mutually deciding treatment regimens, do not necessarily lengthen the overall medical encounter. Furthermore, collaboratively decided treatment plans are associated with decreased rates of hospitalization and emergency room use. Despite the challenges that exist, providers are implored to attempt patient-centered communication.

Big data has helped facilitate asynchronous communication between medical providers, namely through electronic health records which ensure continuity of care, but big data’s real promise lies elsewhere. Using the power of predictive analytics, big data can play an important role in advancing patient-centered health by helping shape tailored wellness programs. The provider-driven, disease-focused approach to health care has, heretofore, impacted the kind of health data that exist: data that are largely focused on patients’ symptoms and diseases. However, diseases do not develop in isolation. Most conditions develop through a complicated interplay of hereditary, environmental, and lifestyle factors. Expanding health data to include social and behavioral data, elicited via a biopsychosocial/patient-centered approach, can help medical providers build better predictive models. By examining comprehensive rather than disease-focused data, providers can, for example, leverage health data to predict which patients will participate in wellness programs, their level of commitment, and their potential for success. This can be done using data mining techniques, like collaborative filtering. In much the same way Amazon makes purchase

recommendations for its users, providers may similarly recommend wellness programs by taking into account patients’ past behavior and health outcomes. Comprehensive data could also be useful for tailoring different types of programs based on patients’ preferences, thereby facilitating increased participation and retention. For example, programs could be customized for patients that go beyond traditional racial, ethnic, or socio-demographic markers and include characteristics such as social media use and shopping habits. By designing analytics aimed at understanding individual patients and not just their diseases, providers may better grasp how to motivate and support the necessary behavioral changes required for improved health.

The International Olympic Committee (IOC), in a consensus meeting on noncommunicable disease prevention, has called for an expansion of health data collected and a subsequent conversion of that data into information providers and patients may use to achieve better health outcomes. Noncommunicable/chronic diseases, such as diabetes and high blood pressure, are largely preventable. These conditions are related to lifestyle choices: too little exercise, an unhealthy diet, smoking, and alcohol abuse. The IOC recommends capturing data from pedometers and sensors in smart phones, which provide details about patients’ physical activity, and combining that with data from interactive smart phone applications (such as calorie counters and food logs) to customize behavior counseling. This approach individualizes not only patient care but also education, prevention, and treatment interventions and advances patient-centered care with respect to information sharing, participation, and collaboration. The IOC also identifies several other potential sources of health data: social medial profiles, electronic medical records, and purchase histories. Collectively, this data can yield a “mass customization” of prevention programs. Given chronic diseases are responsible for 60 percent of deaths and 80 percent of healthcare spending is dedicated to chronic disease management, customizable programs have the potential to save lives and money.

Despite the potential, big data's impact are largely unrealized in patient-centered care efforts. Although merging social, behavioral, and medical data to improve health outcomes has not happened on a widespread basis, there is still a lot that can be done analyzing medical data alone. There is, however, a clear need for computational/analytical tools that can aid providers in recognizing disease patterns, predicting individual patients' susceptibility, and developing personalized interventions. Nitesh Chawla and Darcy Davis propose aggregating and integrating big data derived from millions of electronic health records to uncover patients' similarities and connections with respect to numerous diseases. This makes a proactive medical model possible, as opposed to the current treatment-based approach. Chawla and Davis suggest that leveraging clinically reported symptoms from a multitude of patients, along with their health histories, prescribed treatments, and wellness strategies, can provide a summary report of possible risk factors, underlying causes, and anticipated concomitant conditions for individual patients. They developed an analytical framework called the Collaborative Assessment and Recommendation Engine (CARE), which applies collaborative filtering using inverse frequency and vector similarity to generate predictions based on data from similar patients. The model was validated using a Medicare database of 13 million patients with two million hospital visits over a 4-year period by comparing diagnosis codes, patient histories, and health outcomes. CARE generates a short list that includes high-risk diseases and early warning signs that a patient may develop in the future, enabling a collaborative prevention strategy and better health outcomes. Using this framework, providers can improve the quality of care through prevention and early detection and also advance patient-centered health care.

Data security is a factor that merits discussion. Presently, healthcare systems and individual providers exclusively manage patients' health data. Healthcare systems must comply with security mandates set forth by the Health Insurance Portability and Accountability Act of 1996 (HIPAA). HIPAA demands data servers are firewall and

password protected, and use encrypted data transmission. Information sharing is an important component of patient-centered care. Some proponents of the patient-centered care model advocate transferring control of health data to patients, who may then use and share it as they see fit. Regardless as to who maintains control of health data, storing and electronically transferring that data pose potential security and privacy risks.

Patient-centered care requires collaborative partnerships and wellness strategies that incorporate patients' thoughts, feelings, and preferences. It also requires individualized care, tailored to meet patients' unique needs. Big data facilitates patient-centered/individualized care in several ways. First, it ensures continuity of care and enhanced information sharing through integrated electronic health records. Second, analyzing patterns embedded in big data can help predict disease. APACHE III, for example, is a prognostic program that predicts hospital inpatient mortality. Similar programs help predict the likelihood of heart disease, Alzheimer's, cancer, and digestive disorders. Lastly, big data accrued from not only patients' health records but from their social media profiles, purchase histories, and smartphone applications have the potential to predict enrollment in wellness programs and improve behavioral modification strategies thereby improving health outcomes.

Cross-References

- ▶ [Biomedical Data](#)
- ▶ [Electronic Health Records \(EHR\)](#)
- ▶ [Epidemiology](#)
- ▶ [Health Care Delivery](#)
- ▶ [Health Informatics](#)
- ▶ [HIPAA](#)
- ▶ [Predictive Analytics](#)

Further Reading

- Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(3), 660–665.

- Duffy, T. P. (2011). The Flexner report: 100 years later. *Yale Journal of Biology and Medicine*, 84(3), 269–276.
- Institute of Medicine. (2001). *Crossing the quality chasm*. Washington, DC: National Academies Press.
- Institute for Patient- and Family-Centered Care. *FAQs*. <http://www.ipfcc.org/faq.html>. Accessed Oct 2014.
- Matheson, G., et al. (2013). Prevention and management of non-communicable disease: The IOC consensus statement, Lausanne 2013. *Sports Medicine*, 43, 1075–1088.
- Picker Institute. *Principles of patient-centered care*. http://pickerinstitute.org/about/picker_principles/. Accessed Oct 2014.

PatientsLikeMe

Niccolò Tempini

Department of Sociology, Philosophy and Anthropology and Egenis, Centre for the Study of the Life Sciences, University of Exeter, Exeter, UK

Introduction

PatientsLikeMe is a for-profit organization based in Cambridge, Massachusetts, managing a social media-based health network that supports patients in activities of health data self-reporting and socialization. As of January 2015, the network counts more than 300,000 members and 2,300+ associated conditions and it is one of the most established networks in the health social media space. The web-based system is designed and managed to encourage and enable patients to share data about their health situation and experience.

Business Model

Differently from most prominent social media sites, the network is not ad-supported. Instead, the business model centers on the sale of anonymized data access and medical research services to commercial organizations (mostly pharmaceutical companies). The organization has been partnering with clients, in order to develop patient

communities targeted on a specific disease, or kind of patient experience. In the context of a sponsored project, *PatientsLikeMe* staff develop disease-specific tools required for patient health self-reporting (Patient-reported outcome measures – PROMs) on a web-based platform, then collect and analyze the patient data, and produce research outputs, either commercial research reports or peer-reviewed studies. Research has regarded a wide range of issues, from drug efficacy discovery for neurodegenerative diseases, or symptom distribution across patient populations, to sociopsychological issues like compulsive gambling.

While the network has produced much of its research in occasion of sponsored research projects, this has mostly been discounted from criticism. This because, for its widespread involvement of patients in medical research, *PatientsLikeMe* is often seen as a champion of the so-called participatory turn in medicine, the issue of patient empowerment and more generally of the forces of democratization that several writers argued to be promise of the social web. While sustaining its operations through partnerships with commercial corporations, *PatientsLikeMe* also gathers on the platform a number of patient-activism NGOs. The system provides them customized profiles and communication tools, with which these organizations can try to improve the reach with the patient population of reference, while the network in return gains a prominent position as the center, or enabler, of health community life.

Patient Members

PatientsLikeMe attracts patient members because the system is designed to allow patients to find others and socialize. This can be particularly useful for patients of rare, chronic, or life-changing diseases: patient experiences for which an individual might feel helpful to learn from the experience of others, whom however might be not easy to find through traditional, “offline” socialization opportunities. The system is also designed to enable self-tracking of a number of health dimensions. The patients record both structured data,

about diagnoses, treatments, symptoms, disease-specific patient-reported questionnaires (PROs), or results of specific lab test, and semi-structured or unstructured data, in the form of comments, messages, and forum posts. All of these data are at the disposal of the researchers that have access to the data. A paradigmatic characteristic of *PatientsLikeMe* as social media research network is that the researchers do not learn about the patients in any other way than through the data that the patients share.

Big Data and *PatientsLikeMe*

As such, it is the approach to data and to research that defines *PatientsLikeMe* as a representative “Big Data” research network – one that, however, does not manage staggeringly huge quantities of data nor employs extremely complex technological solutions for data storage and analysis. *PatientsLikeMe* is a big data enterprise because, first, it approaches medical research through an open (to data sharing by anyone and about user-defined medical entities), distributed (relative to availability of a broadband connection, from anywhere and at anytime), and data-based (data are all that is transacted between the participating parties) research approach. Second, the data used by *PatientsLikeMe* researchers are highly varied (including social data, social media user-generated content, browsing session data, and most importantly structured and unstructured health data) and relatively fast, as they are updated, parsed, and visualized dynamically in real time through the website or other data-management technologies. The research process involves practices of pattern detection, analysis of correlations, and investigation of hypotheses through regression and other statistical techniques.

The vision of scientific discovery that is underlying the *PatientsLikeMe* project is one based on the assumption that given a broad enough base of users and a granular, frequent and longitudinal exercise of data collection, new, small patterns ought to emerge from the data and invite further investigation and explanation. This assumption

implies that for medical matters to be discovered further, the development of an open, distributed and data-based socio-technical system that is more sensitive to their forms and differences is a necessary step. But also, the hope is that important lessons can be learned by opening the medical framework to measure and represent a broader collection of entities and events than traditional, profession-bound medical practice accepted. The *PatientsLikeMe* database includes symptoms and medical entities as described in the terms used by the patients themselves. This involves sensitive and innovative processes of translation from the patient language to expert terminology. Questions about the epistemological consequence of the translation of the patient voice (until now a neglected form of medical information) over data fields and categories, and the associated concerns about reliability of patient-generated data, cannot have a simple answer. In any case, from a practice-based point of view these data are nonetheless being mobilized for research through innovative technological solutions for coordinating the patient user-base. The data can then be analyzed in multiple ways, all of which include the use of computational resources and databases – given the digital nature of the data.

As ethnographic research of the organization has pointed out (see further readings section, below), social media companies that try to develop knowledge from the aggregation and analysis of the data contributed by their patients are involved in complex efforts to “cultivate” the information lying in the database – as they have to come to grips with the dynamics and trade-offs that are specific to understanding health through social media. Social media organizations try to develop meaningful and actionable information from their database by trying to make data structures more precise in differentiating between phenomena and reporting about them in data records, and make the system easier and flexible in use in order to generate more data. Often these demands work at cross-purposes. The development of social media for producing new knowledge through distributed publics involves the engineering of social environment where sociality and information production are inextricably

intertwined. Users need to be steered towards information-productive behaviors as they engage in social interaction of sorts, for information is the worth upon which social media businesses depend. In this respect, it has been argued that *PatientsLikeMe* is representative of the construction of sociality that takes place in all social media sites, where social interaction unfolds along the paths that the technology continuously and dynamically draws based on the data that the users are sharing.

As such, many see *PatientsLikeMe* as incarnating an important dimension of the much-expected revolution of personalized medicine. Improvements in healthcare will not be limited to a capillary application of genetic sequencing and other micro and molecular biology tests that try to open up the workings of individual human physiology at unprecedented scale, instead the information produced by these tests will often be related with the information about the subjective patient experience and expectations that new information technology capabilities are increasingly making possible.

Other Issues

Much of the public debate about the *PatientsLikeMe* network involves issues of privacy and confidentiality of the patient users. The network is a “walled garden,” with patient profiles remaining inaccessible to unregistered users by default. However, once logged in, every user can browse all patient profiles and forum conversations. In more than one occasion, unauthorized intruders (including journalists and academics) were detected and found screen-scraping data from the website. Despite the organization employing state-of-the-art techniques to protect patient data from unauthorized exporting, any sensitive data shared on a website remains at a risk, given the widespread belief – and public record on other websites and systems – that skilled intruders could always execute similar exploits unnoticed. Patients can have a lot to be concerned about, especially if they have conditions with a social stigma or if they shared explicit political or

personal views in the virtual comfort of a forum room. In this respect, even if the commercial projects that the organization has undertaken with industry partners implied the exchange of user data that had been pseudonymised before being handed over, the limits of user profile anonymization are well known. In the case of profiles of patients living with rare diseases, which are a consistent portion of the users in *PatientsLikeMe*, it can arguably be not too difficult to reidentify individuals, upon determined effort. These issues of privacy and confidentiality remain a highly sensitive topic as society does not dispose of standard and reliable solutions against the various forms that data misuse can take. As both news and scholars have often reported, the malleability of digital data makes it impossible to stop the diffusion of sensitive data once that function creep happens.

Moreover, as it is often discussed in the social media and big data public debate, data networks increasingly put pressure on the notion of informed consent as an ethically sufficient device for conducting research with user and patient data. The need for moral frameworks of operation that overperform over strict compliance with law has often been called for, and recently by the report on data in biomedical research by the Nuffield Council for Bioethics. In the report, *PatientsLikeMe* was held as a paramount example of new kinds of research networks that rely on extensive patient involvement and social (medical) data – these networks are often dubbed as citizen science or participatory research.

On another note, some have argued that *PatientsLikeMe*, as many other prominent social media organizations, has been exploiting the rhetoric of sharing (one’s life with a network and its members) to encourage data-productive behaviors. The business model of the network is built around a traditional, proprietary model of data ownership. The network facilitates the data flow inbound and makes it less easy for the data to flow outbound, controlling their commercial application. In this respect, we must notice that the current practice in social media management in general is often characterized by data sharing evangelism by the managing organization, which

at the same time requires monopoly of the most important data resources that the network generates. In the general public debate, this kind of social media business model has been linked as a factor contributing to the erosion of user privacy.

On a different level, one can notice how the kind of patient-reported data collection and medical research that the network makes possible to perform is a much cheaper and under many respects more efficient model than what the professional-laden institutions such as the clinical research hospital, with their specific work loci and customs, could put in place. This way of organising the collection of valuable data operates by including large amounts of end users who are not remunerated. Despite this, running and organizing such an enterprise is expensive and labor-intensive and as such, critical analysis of this kind of “crowdsourcing” enterprise needs to look beyond the more superficial issue of the absence of a contract to sanction the exchange of a monetary reward for distributed, small task performances. One connected problem in this respect is that since data express their value only when they are re-situated through use, no data have a distinct, intrinsic value upon generation; not all data generated will ever be equal.

Finally, the affluence of medical data that this network makes available can have important consequences on therapy or lifestyle decisions that a patient might take. Sure, patients can make up their mind and take critical decisions without appropriate consultation at any time, as they have always done. Nonetheless, the sheer amount of information that networks such as *PatientsLikeMe* or search engines such as Google make available at a click’s distance is without antecedents and what this implies for healthcare must still be fully understood. Autonomous decisions by the patients do not necessarily happen for the worst. As healthcare often falls short of providing appropriate information and counseling, especially about everything that is not strictly therapeutic, patients can eventually devise improved courses of action, through a consultation of appropriate information-rich web resources. At the same time, risks and harms are not fully appreciated and there

is a pressing need to understand more on the consequences of these networks for individual health and the future of healthcare and health research.

There are other issues besides these more evident and established topics of discussion. As it has been pointed out, questions of knowledge translation (from the patient vocabulary to the clinical-professional) remain open, and unclear is also the capacity of these distributed and participative networks to consistently represent and organize the patient populations that they are deemed to serve, as the involvement of patients is however limited and relative to specific tasks, most often of data-productive character. The afore-mentioned issues are not exhaustive nor exhausted in this essay. They require in-depth treatment; with this introduction the aim has been to give a few coordinates on how to think about the subject.

Further Reading

- Angwin, J. (2014). *Dragnet nation: A quest for privacy, security, and freedom in a world of relentless surveillance*. New York: Henry Holt and Company.
- Arnott-Smith, C., & Wicks, P. (2008). PatientsLikeMe: Consumer health vocabulary as a folksonomy. *American Medical Informatics Association Annual Symposium Proceedings*, 2008, 682–686.
- Kallinikos, J., & Tempini, N. (2014). Patient data as medical facts: Social media practices as a foundation for medical knowledge creation. *Information Systems Research*, 25, 817–833. <https://doi.org/10.1287/isre.2014.0544>.
- Lunshof, J. E., Church, G. M., & Prainsack, B. (2014). Raw personal data: Providing access. *Science*, 343, 373–374. <https://doi.org/10.1126/science.1249382>.
- Prainsack, B. (2013). Let’s get real about virtual: Online health is here to stay. *Genetical Research*, 95, 111–113. <https://doi.org/10.1017/S001667231300013X>.
- Richards, M., Anderson, R., Hinde, S., Kaye, J., Lucassen, A., Matthews, P., Parker, M., Shotter, M., Watts, G., Wallace, S., & Wise, J. (2015). *The collection, linking and use of data in biomedical research and health care: Ethical issues*. London: Nuffield Council on Bioethics.
- Tempini, N. (2014). *Governing social media: Organising information production and sociality through open, distributed and data-based systems* (Doctoral dissertation). School of Economics and Political Science, London.
- Tempini, N. (2015). *Governing PatientsLikeMe: Information production and research through an open,*

- distributed and data-based social media network. *The Information Society*, 31, 193–211.
- Wicks, P., Vaughan, T. E., Massagli, M. P., & Heywood, J. (2011). Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, 29, 411–414. <https://doi.org/10.1038/nbt.1837>.
- Wyatt, S., Harris, A., Adams, S., & Kelly, S. E. (2013). Illness online: Self-reported data and questions of trust in medical and social research. *Theory Culture & Society*, 30, 131–150. <https://doi.org/10.1177/0263276413485900>.
- Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30, 75–89.

Pattern Recognition

► Financial Data and Trend Prediction

Persistent Identifiers (PIDs) for Cultural Heritage

Jong-On Hahm

Department of Chemistry, Georgetown University, Washington, DC, USA

A persistent identifier (PID) is a long-lasting reference to a digital resource. Examples include digital object identifiers (DOIs) for publications and datasets, and ORCID iDs for individual authors. A PID can provide access to large amounts of data and metadata about an object, offering a diverse array of information previously unavailable to the public.

Cultural heritage science, specifically the conservation and characterization of artworks and antiquities, would greatly benefit from the establishment of a system of persistent identifiers. Cultural heritage objects are treasured as historical and cultural assets that can iconify the national identity of many societies. They are also resources for education, drivers of economic activity, and represent significant financial assets. A system of persistent identifiers for cultural heritage could

serve as a source of big data for a number of economic sectors.

One project that could demonstrate the potential of persistent identifiers as a target of big data was recently launched in the United Kingdom (UK). Towards a National Collection is a multi-year, \$25.4 M project to develop a virtual national collection of the cultural heritage assets held in the UK's museums, galleries, libraries, and archives. Part of the project is an initiative to establish a PID system as a “research infrastructure,” a data resource available to enable data discovery and access. The goal of the project is to spur research and innovation, as well as economic and social benefits. This national investment by UK Research and Innovation ([UKRI.org](https://www.ukri.org/)), a public agency, was undertaken in recognition of the economic power of cultural heritage tourism, and to set global standards for cultural heritage research.

Given the sweeping national level of this effort, it is quite likely that once established, persistent identifiers will spread from the public to the private sector. Buyers of art, particularly of high-end artworks, may want their purchases to be accompanied by the full panoply of information available via a persistent identifier. Insurance companies are even more likely to require that artworks have persistent identifiers, to be able to examine full information and documentation on the object to be insured. As such, the adoption of persistent identifiers for private collections of artworks and antiquities has the potential to fundamentally and dramatically transform the art and insurance markets.

In 2019, the global art market was a ~\$64.1 billion enterprise of which 42% of total sale values were objects priced more than \$1 million. According to the US Department of Justice, art crime is the third highest-grossing criminal trade. It is also one of the least prosecuted, primarily because data on art objects are scarce. The establishment of a persistent identifier system could be a disruptive force in art crime, a global enterprise of which the vast majority of sales and transactions goes undetected.

Further Reading

Art scandal threatens to expose mass fraud in global art market. <https://www.cnn.com/2015/03/13/art-scandal-threatens-to-expose-mass-fraud-in-global-art-market.html>.

Persistent Identifiers as IRO Infrastructure. <https://bl.iro.bl.uk/work/ns/14d713d7-72d3-4f60-8583-91669758ab41>.

Protecting Cultural Heritage from Art Theft. <https://leb.fbi.gov/articles/featured-articles/protecting-cultural-heritage-from-art-theft-international-challenge-local-opportunity>.

The Art Market 2020. <https://theartmarket.foleon.com/2020/artbasel/the-global-art-market>.

Towards a National Collection. <https://tanc-ahrc.github.io/HeritagePIDs/index.html>.

Personally Identifiable Information

► Anonymization Techniques

Pharmaceutical Industry

Janelle Applequist

The Zimmerman School of Advertising and Mass Communications, University of South Florida, Tampa, FL, USA

Globally, the pharmaceutical industry is worth more than \$1 trillion, encompassing one of the world's most profitable industries, focusing on the development, production, and marketing of prescription drugs for use by patients. Over one-third of the pharmaceutical industry is controlled by just ten companies, with six of these companies in the United States alone. The World Health Organization has reported an inherent conflict of interest between the pharmaceutical industry's business goals and the medical needs of the public, attributable to the fact that twice the amount is spent on promotional spending (including advertisements, marketing, and sales representation) than is on the research and development for future prescription drugs needed for public health

efforts. The average pharmaceutical company in the United States sees a profit of greater than \$10 billion annually, while pharmaceutical companies contribute 50 times more spending on promoting and advertising for their own products than spending on public health information initiatives.

Big data can be described as the collection, manipulation, and analysis of massive amounts of data – and the decisions made from that analysis. Having the ability to be described as both a problem and an opportunity, big data and its techniques are continuing to be utilized in business by thousands of major institutions. The sector of health care is not immune to massive data collection efforts, and pharmaceuticals in particular comprise an industry that relies on aggregating information.

Literature on data mining in the pharmaceutical industry generally points to a disagreement regarding the intended use of health-care information. On the one hand, historically, data mining techniques have proved useful for the research and development (R&D) of current and future prescription drugs. Alternatively, continuing consumerist discourses in health care that have positioned the pharmaceutical industry as a massive and successful corporate entity have acknowledged how this data is used to increase business sales, potentially at the cost of patient confidentiality and trust.

History of Data Mining Used for Pharmaceutical R&D

Proponents of data mining in the pharmaceutical industry have cited its ability to aid in: organizing information pertaining to genes, proteins, diseases, organisms, and chemical substances, allowing predictive models to be built for analyzing the stages of drug development; keeping track of adverse effects of drugs in a neural network during clinical trial stages; listing warnings and known reactions reported during the post-drug production stage; forecasting new drugs needed in the marketplace; providing inventory control and supply chain management information; and managing inventories. Data mining was first used

in the pharmaceutical industry as early as the 1960s alongside the increase in prescription drug patenting. With over 1,000 drug patents a year being introduced at that time, data collection assisted pharmaceutical scientists in keeping up with patents being proposed. At this time, information was collected and published in an editorial-style bulletin categorized according to areas of interest in an effort to make relevant issues for scientists easier to navigate. Early in the 1980s, technologies allowed biological sequences to be identified and stored, such as the Human Genome Project, which led to the increased use and publishing of databanks. Occurring alongside the popularity of personal computer usage, bioinformatics was born, which allowed biological sequence data to be used for discovering and studying new prescription drug targets. Ten years later, in the 1990s, microarray technology developed, posing a problem for data collection, as this technology permitted the simultaneous measurement of large numbers of genes and collection of experimental data on a large scale. As the ability to sequence a genome occurred in the 2000s, the ability to manage large levels of raw data was still maturing, creating a continued problem for data mining in the pharmaceutical industry. As the challenges presented for data mining in relation to R&D have continued to increase since the 1990s, the opportunities for data mining in order to increase prescription drug sales have steadily grown.

Data Mining in the Pharmaceutical Industry as a Form of Controversy

Since the early 1990s, health-care information companies have been purchasing the electronic records of prescriptions from pharmacies and other data collection resources in order to strategically link this information with specific physicians.

Prescription tracking refers to the collection of data from prescriptions as they are filled at pharmacies. When a prescription gets filled, data miners are able to collect: the name of the drug, the date of the prescription, and the name or

licensing number of the prescribing physician. Yet, it is simple for the prescription drug industry to identify specific physicians through protocol in place by the American Medical Association (AMA). The AMA has a "Physician Masterfile" that includes all US physicians, whether or not they belong to the AMA, and this file allows the physician licensing numbers collected by data miners to be connected to a name. Information distribution companies (such as IMS Health, Dendrite, Verispan, Wolters Kluwer, etc.) purchase records from pharmacies. What many consumers do not realize is that most pharmacies have these records for sale and are able to do so legally by not including patient names and only providing a physician's state licensing number and/or name. While pharmacies cannot release a patient's name, they can provide data miners with a patient's age, sex, geographic location, medical conditions, hospitalizations, laboratory tests, insurance copays, and medication use. This has caused a significant area of concern on behalf of patients, as it not only may increase instances of prescription detailing, but it may compromise the interests of patients. Data miners do not have access to patient names when collected prescription data; however, data miners assign unique numbers to individuals so that future prescriptions for the patient can be tracked and analyzed together. This means that data miners can determine: how long a patient remains on a drug, whether the drug treatment is continued, and which new drugs become prescribed for the patient.

As information concerning a patient's health is highly sensitive, data mining techniques used by the pharmaceutical industry have perpetuated the notion that personal information carries a substantial economic value. By data mining companies paying pharmacies to extract prescription drug information, the relationships between patients and their physicians and/or pharmacists is being exploited. The American Medical Association (AMA) established the Physician Data Restriction Program in 2006, giving any physician the opportunity to opt out from data mining initiatives. To date, no such program for patients exists that would give them the opportunity to have their records removed from data collection procedures

and subsequent analyses. Three states have enacted statutes that do not permit data mining of prescription records. The Prescription Confidentiality Act of 2006 in New Hampshire was the first state to decide that prescription information could not be sold or used for any advertising, marketing, or promotional purposes. However, if the information is de-identified, meaning that the physician and patient names cannot be accessed, then the data can be aggregated by geographical region or zip code, meaning that data mining companies could still provide an overall, more generalized report for small geographic areas but could not target specific physicians. Maine and Vermont have statutes that limit the presence of data mining. Physicians in Maine can register with the state to prevent data mining companies from obtaining their prescribing records. Data miners in Vermont must obtain consent from the physician for which they are analyzing prior to using “prescriber-identifiable” information for marketing or promotional purposes.

The number one customer for information distribution companies is the pharmaceutical industry, which purchases the prescribing data to identify the highest prescribers and also to track the effects of their promotional efforts. Physicians are given a value, a ranking from one to ten, which identifies how often they prescribe drugs. A sales training guide for Merck even states that this value issued to identify which products are currently in favor with the physician in order to develop a strategy to change those prescriptions into Merck prescriptions. The empirical evidence provided by information distribution companies offers a glimpse into the personality, behaviors, and beliefs of a physician, which is why these numbers are so valued by the drug industry.

By collecting and analyzing this data, pharmaceutical sales representatives are able to better target their marketing activities toward physicians. For example, as a result of data mining in the pharmaceutical industry, pharmaceutical sales representatives could: determine which physicians are already prescribing specific drugs in order to reinforce already-existent preferences, or, could learn when a physician switches from a drug to a competing drug, so that the

representative can attempt to encourage the physician to switch back to the original prescription.

The Future of Data Mining in the Pharmaceutical Industry

As of 2013, only 18% of pharmaceutical companies work directly with social media to promote their prescription drugs, but this number is expected to increase substantially in the next year. As more individuals tweet about their medical concerns, symptoms, the drugs they take, and respective side effects, pharmaceutical companies have noticed that social media has become an integrated part of personalized medicine for individuals. Pharmaceutical companies are already in the process of hiring data miners to collect and analyze various forms of public social media in an effort to: discover unmet needs, recognize new adverse events, and determine what types of drugs consumers would like to enter the market.

Based on the history of data mining used by pharmaceutical corporations, it is evident that the lucrative nature of prescription drugs serves as a catalyst for data collection and analysis. By having the ability to generalize what should be very private information about patients for the prescription drug industry, the use of data allows prescription drugs to make more profit than ever, as individual information can be commoditized to benefit the bottom line of a corporation. Although there are evident problems associated with prescription drug data mining, the US Supreme Court has continued to recognize that the pharmaceutical industry has a first amendment right to advertise and solicit clients for goods and future services. The Court has argued that legal safeguards, such as the Health Information Portability and Accountability Act (HIPAA), are put in place to combat the very concerns posed by practices such as pharmaceutical industry data mining. Additionally, the Court has found that by stripping pharmaceutical records of patient information that could lead to personal identification (e.g., name, address, etc.), patients have their confidentiality adequately protected. The law, therefore, leaves it to the discretion of the physician to decide

whether they will associate with pharmaceutical sales representatives and various data collection procedures.

An ongoing element to address in analyzing the pharmaceutical industry's use of data mining techniques will be the level of transparency used with patients while utilizing the information collected. Research shows that the majority of patients in the United States are not only unfamiliar with data mining use by the pharmaceutical industry, but that they are against any personal information (e.g., prescription usage information and personal diagnoses) being sold and shared with outside entities, namely, corporations. As health care continues to change in the United States, it will be important for patients to understand the ways in which their personal information is being shared and used, in an effort to increase national understandings of how privacy laws are connected to the pharmaceutical industry.

Cross-References

- [Electronic Health Records \(EHR\)](#)
- [Health Care Delivery](#)
- [Patient Records](#)
- [Privacy](#)

Further Reading

- Altan, S., et al. (2010). Statistical considerations in design space development. *Pharmaceutical Technology*, 34(7), 66–70.
- Fugh-Berman, A. (2008). Prescription tracking and public health. *Journal of General Internal Medicine*, 23(8), 1277–1280.
- Greene, J. A. (2007). Pharmaceutical marketing research and the prescribing physician. *Annals of Internal Medicine*, 146(10), 742–747.
- Klocke, J. L. (2008). Comment: Prescription records for sale: Privacy and free speech issues arising from the sale of de-identified medical data. *Idaho Law Review*, 44(2), 511–536.
- Orentlicher, D. (2010). Prescription data mining and the protection of patients' interests. *The Journal of Law, Medicine & Ethics*, 38(1), 74–84.
- Steinbrook, R. (2006). For sale: Physicians' prescribing data. *The New England Journal of Medicine*, 354(26), 2745–2747.

Wang, J., et al. (2011). Applications of data mining in pharmaceutical industry. *The Journal of Management and Engineering Integration*, 4(1), 120–128.

White paper: Big Data and the needs of the Pharmaceutical Industry. (2013). Philadelphia: Thomson Reuters.

World Health Organization. (2013). *Pharmaceutical Industry*. Retrieved online from <http://www.who.int/trade/glossary/story073/en/>.

Policy

- [Regulation](#)

Policy Analytics

Laurie A. Schintler

George Mason University, Fairfax, VA, USA

Overview

Over the last half century, the policymaking process has undergone a digital transformation (Pencheva et al. 2020). Information technology such as computers and the Internet – artifacts of the “digital revolution” – helped usher in data-driven public policy analysis and decision-making in the 1980s (Gil-Garcia et al. 2018). Now big data, coupled with new and advancing computational tools and analytics (e.g., machine learning), are digitalizing the process even further. While the origins of the term are murky, policy analytics encapsulates this changing context, referring specifically to the use of big data resources and tools for policy analysis (Daniell et al. 2016). Although policy analytics can benefit the policymaking process in various ways, it also comes with a set of issues, challenges, and downsides that must be managed simultaneously.

Prospects and Potentialities

Policymaking involves developing, analyzing, evaluating, and implementing laws, regulations,

and other courses of action to solve real-world problems for improving societal welfare. The policy cycle is a framework for understanding this process and the “complex strategic activities, actors, and drivers” it affects and is affected by (Pencheva et al. 2020). Critical steps in this cycle include:

1. Problem identification and agenda setting
2. Development of possible policy options (or policy instruments)
3. Evaluation of the feasibility and impact of each policy option
4. Selection and implementation of a policy or set of guidelines

There is also often an ongoing assessment of policies and their impacts after they have been implemented (i.e., ex-postevaluation), which may, in turn, result in the modification or termination of policies.

Data and methods have long played a critical role in all phases of the policy life cycle. In this regard, various types and forms of qualitative and quantitative data (e.g., census records, household surveys), along with models and tools (e.g., cost-benefit analysis, statistical inference, and mathematical optimization), are used for analyzing and assessing policy problems and their potential solutions (Daniell et al. 2016). Big data and data-driven computational and analytical tools (e.g., machine learning) provide a new “toolbox” for policymakers and policy analysts, which can help address the growing complexities of the policymaking process while overcoming the limitations of conventional methods and data sources for policy analysis.

First, big data provide a rich means for identifying, characterizing, and tracking problems for which there may be a need for policy solutions. Indeed, such tasks are fraught with a growing number of complications and challenges, as public issues and public policies have become increasingly dynamic, interconnected, and unpredictable (Renteria and Gil-Garcia 2017). In this regard, conventional sources of data (e.g., government censuses) tend to fall short, especially given the information is described in fixed and

aggregate forms, spatially and temporally. Accordingly, such data lack the level of resolution required to understand the details and nuances of public problems, such as how particular individuals, neighborhoods, and groups are negatively impacted by dynamic situations and circumstances. Big data, such as that produced by video surveillance cameras, the Internet of Things (IoT), mobile phones, and social media, provide the granularity to address such gaps.

Second, data-driven analytics such as deep neural learning – a powerful form of machine learning that attempts to mimic how the human brain processes information – has enormous potential for policy analysis. Specifically, such approaches enable insight to be gleaned from massive amounts of streaming data in a capacity not possible with traditional models and frameworks. Moreover, supervised machine learning techniques for prediction and classification can help anticipate trends and evaluate policy options on the fly. They also give policymakers the ability to test potential solutions in advance. “Nowcasting,” an approach developed in the field of economics, enables the evaluation of policies in the present, the imminent future, and the recent past (Bańbura et al. 2010). Such methods can supplement and inform models used for longer-term forecasting.

Third, big data produced by crowdsourcing mechanisms and platforms provide a valuable resource for addressing the difficulties associated with agenda setting, problem identification, and policy prioritization (Schintler and Kulkarni 2014). Such activities pose an array of challenges. One issue is that the policymaking process involves multiple stakeholders, each of which has its own set of values, objectives, expectations, interests, preferences, and motivations. Complicating matters is that positions on “hot-button” policy issues, such as climate change and vaccinations, have become increasingly polarized and politicized. While surveys and interviews provide a means for sensing the opinions, attitudes, and needs of citizens and other stakeholders, they are costly and time-consuming to implement. Moreover, they do not allow for real-time situational awareness and intelligence. Crowdsensing data,

combined with tools such as sentiment analysis, provide a potential means for understanding, tracking, and accounting for ongoing and fluctuating views on policy problems and public policy solutions. Thus, as a lever for increasing participation in the policy cycle, crowdsensed big data can promote social and civic empowerment, ultimately engendering trust within and between stakeholder groups (Brabham 2009).

Downsides, Dilemmas, and Challenges

Despite the actual and potential benefits of big data and data-driven methods for policy analysis, i.e., policy analytics, the public sector has yet to make systematic and aggressive use of such tools, resources, and approaches (Daniell et al. 2016; Sun and Medaglia 2019). While robust methods, techniques, and platforms for big data have been developed for business (i.e., business intelligence), they cannot (and should not) be transferred to a public policy context. One significant issue is that framings, interests, biases, and motivations – and values – of public and private entities tend to be incongruent (Sun and Medaglia 2019). Whereas companies generally strive to maximize rate profit and rate-of-return on investment, the government is more concerned with equitably allocating public resources to promote societal well-being (Daniell et al. 2016) (Of course, there are some exceptions, e.g., “socially-conscious” corporations or corrupt governments). As values get embedded into the architecture and design of computational models and drive the selection and use of data in the first place, the blind application of business analytics to public policy can have dangerous consequences. More to the point, while the use of business intelligence for policy analysis may yield efficient and cost-saving policy solutions, it may come at the expense of broader societal interests, such as human rights and social justice. On top of all this, there are technical and ethical issues and considerations that come into play in applying big data and data-driven methods in the public sphere, which create an additional set of barriers to

their use and application. One challenge in this regard is integrating traditional sources of data (e.g., census records) with big data, especially given they tend to have different levels of resolution and coverage. Issues related to privacy, data integrity, data provenance, and algorithmic bias and discrimination complicate matters further (Schintler 2020; Schintler and Fischer 2018).

Conclusion

In sum, while the use of big data and data-driven methods for policy analysis, i.e., policy analytics, can improve the efficiency and effectiveness of the policymaking process in various ways, it also comes with an array of downsides and dilemmas, as highlighted. Thus, a grand challenge is balancing the need for “robust and convincing analysis” with the need to satisfy public expectations about the transparency, fairness, and integrity of the policy process and its outcomes (Daniell et al. 2016). In this regard, public policy itself has a crucial role to play.

Cross-References

- ▶ [Business Intelligence Analytics](#)
- ▶ [Crowdsourcing](#)
- ▶ [Ethics](#)
- ▶ [Governance](#)

Further Reading

- Bañbura, M., Giannone, D., & Reichlin, L. (2010). *Nowcasting*, ECB working paper, no. 1275. Frankfurt a. M.: European Central Bank (ECB).
- Brabham, D. C. (2009). Crowdsourcing the public participation process for planning projects. *Planning Theory*, 8(3), 242–262.
- Daniell, K. A., Morton, A., & Insua, D. R. (2016). Policy analysis and policy analytics. *Annals of Operations Research*, 236(1), 1–13.
- Gil-Garcia, J. R., Pardo, T. A., & Luna-Reyes, L. F. (2018). Policy analytics: Definitions, components, methods, and illustrative examples. In *Policy analytics, modeling, and informatics* (pp. 1–16). Cham: Springer.

- Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2020). Big data and AI—A transformational shift for government: So, what next for research? *Public Policy and Administration*, 35(1), 24–44.
- Renteria, C., & Gil-Garcia, J. R. (2017). A systematic literature review of the relationships between policy analysis and information technologies: Understanding and integrating multiple conceptualizations. In *International conference on electronic participation* (pp. 112–124). Cham: Springer.
- Schintler, L.A. (2020). Regional policy analysis in the era of spatial big data. In *Development studies in regional science* (pp. 93–109). Singapore: Springer.
- Schintler, L. A., & Kulkarni, R. (2014). Big data for policy analysis: The good, the bad, and the ugly. *Review of Policy Research*, 31(4), 343–348.
- Schintler, L.A., & Fischer, M. M. (2018). Big data and regional science: Opportunities, challenges and directions for future research (Working Papers in Regional Science). WU Vienna University of Economics and Business, Vienna. https://epub.wu.ac.at/6122/1/Fischer_etal_2018_Big-data.pdf.
- Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, 36(2), 368–383.

Political Science

Marco Morini

Dipartimento di Comunicazione e Ricerca Sociale, Università degli Studi “La Sapienza”, Roma, Italy

Political science is a social science discipline focused on the study of the state, nation, government, and public policies. As a separate field, it is a relatively late arrival, and it is commonly divided into distinct sub-disciplines which together constitute the field: political theory, comparative politics, public administration, and political methodology. Although it seems that political science has been using machine learning methods for decades, nowadays political scientists are encountering larger datasets with increasingly complex structures and are using innovative new big data techniques and methods to collect data and test hypotheses.

Political science deals extensively with the allocation and transfer of power in decision-making, the roles and systems of governance including governments, international organizations, political behavior, and public policies. It is methodologically diverse and employs many methods originating in social science research. Approaches include positivism, rational choice theory, behavioralism, structuralism, post-structuralism, realism, institutionalism, and pluralism.

Although it was codified in the nineteenth century, political science originated in Ancient Greece with the works of Plato and Aristotle. During the Italian Renaissance, Florentine Philosopher Niccolò Machiavelli established the emphasis of modern political science on direct empirical observation of political institutions and actors. Later, the expansion of the scientific paradigm during the Enlightenment further pushed the study of politics beyond normative determinations.

Because political science is essentially a study of human behavior, in all sides of politics, observations in controlled environments are often challenging to reproduce or duplicate, though experimental methods are increasingly common. Because of this, political scientists have historically observed political elites, institutions, and individual or group behavior in order to identify patterns, draw generalizations, and build social and political theories.

Like all social sciences, political science faces the difficulty of observing human actors that can only be partially observed and who have the capacity for making conscious choices. Despite the complexities, contemporary political science has progressed by adopting a variety of methods and theoretical approaches to understanding politics, and methodological pluralism is a defining feature of contemporary political science. Often in contrast with national media, political science scholars seek to compile long-term data and research on the impact of political issues, producing in-depth articles and breaking down the issues.

Several scholars have long been using machine learning methods to develop and analyze relatively large datasets of political events, such as

using multidimensional scaling methods to study roll-call votes from the US Congress. Since decades, therefore, mainstream political methodology has already dealt with the exact attributes that characterize big data – the use of computationally intensive techniques to analyze what to social scientists are large and complex datasets. In 1998, Yale Professors Don Green and Alan Gerber conducted the first randomized controlled trial in modern political science, assigning New Haven voters to receive non-partisan election reminders by mail, phone, or in-person visit from a canvasser and measuring which group saw the greatest increase in turnout. The subsequent wave-of-field experiments by Green, Gerber, and their followers focused on mobilization, testing competing modes of contact and get-out-the-vote language to see which were most successful.

But while there has been this long tradition in political science for big data like research, political scientists are now using innovative new big data techniques and methods to collect data and test hypotheses. They employ automated analytical methods data to create new knowledge from the unstructured and overwhelming amount of data streaming in from a variety of sources. As field experiments are an important methodology that many social scientists use to test many different behavioral theories, now large-scale field experiments can be accomplished at low cost.

Political scientists are seeing interesting new research opportunities with social media data, with large aggregations of field experiment and polling data, and with other large-scale datasets that just a few years ago could not be easily analyzed with available computational resources. Particularly, recent advances in text mining, automatic coding, and analysis are bringing major changes in two interrelated research subfields: social media and politics and election campaigns. Social media analytics and tools such as Twitter Political Index – that measures Twitter users' sentiments about candidates – allow researchers to track posts of candidates and to study social media habits of politicians and governments. Scholars can now gather, manage, and analyze huge amounts of data. On the other hand, recent

election cycles showed how campaigners count on big data in order to win elections.

The most well-known example is how the Democratic National Committee leveraged big data analytics to better understand and predict voter behavior in the 2012 US elections. The Obama campaign used data analytics and the experimental method to assemble a winning coalition vote by vote. In doing so, it overturned the long dominance of TV advertising in US politics and created something new in the world: a national campaign run like a local ward election, where the interests of individual voters were known and addressed.

The 2012 Obama's campaign used big data to rally individual voters. His approach amounted to a decisive break with twentieth-century tools for tracking public opinion, which consisted of identifying small samples that could be treated as representative of the whole. The electorate could be seen as a collection of individual citizens who could each be measured and assessed on their own terms. This campaign became celebrated for its use of technology – much of it developed by an unusual team of coders and engineers – that redefined how individuals could use the Web, social media, and smartphones to participate in the political process.

Cross-References

- [Curriculum, Higher Education, Humanities](#)
- [Data Mining](#)
- [Social Sciences](#)

Further Reading

- Issenberg, S. (2012). How President Obama's campaign used big data to rally individual voters. <http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/>. Accessed 28 May 2014.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. London: Eamon Dolan/Mariner Books.
- McGann, A. (2006). *The logic of democracy*. Ann Arbor: University of Michigan Press.

Pollution, Air

Zerrin Savaşan

Department of International Relations,
Sub-Department of International Law, Selçuk
University, Konya, Turkey

The air contains many different substances, gases, aerosols, particulate matter, trace metals, and a variety of other compounds. If those are not at the same concentration and change in space, and over time to an extent that the air quality deteriorates, some contaminants or pollutant substances exist in the air. The release of these air pollutants causes harmful effects to both environment and humans, to all organisms. This is regarded as air pollution.

The air is a common/shared resource of all human beings. After released, air pollutants can be carried by natural events like winds, rains, and so on. So, some pollutants, e.g., lead or chloroform, often contaminate more than one environmental occasions, so, many air pollutants can also be water or land pollutants. They can combine with other pollutants and thus can undergo chemical transformations, and then they can be eventually deposited on different locations. Their effects can emerge in different locations far from their main resources. Thus, they can detrimentally affect upon all organisms on local or regional scales and also upon the climate on global scale.

Hence, concern for air pollution and its influences on the earth and efforts to prevent/and to mitigate it have increased greatly in global scale. However, today, it still stands as one of the primary challenges that should be addressed globally on the basis of international cooperation. So, it becomes necessary to promote the widespread understanding on air pollution, its pollutants, sources, and impacts.

Sources of Air Pollution

The air pollutants can be produced from natural-based reasons (e.g., fires from burning vegetation,

forest fires, volcanic eruptions, etc.) or anthropogenic (human-caused) reasons. When outdoor pollution – referring to the pollutants found in outdoors – is thought, smokestacks of industrial plants can be given as an example of human-made ones. However, natural processes also produce outdoor air pollution, e.g., volcanic eruptions. The main causes of indoor air pollution, on the other hand, again raise basically from human-driven reasons, e.g., technologies used for cooking, heating, and lighting. Nonetheless, again there are also natural indoor air pollutants, like radon, and chemical pollutants from building materials and cleaning products.

Among those, human-based reasons, specifically after industrialization, have produced a variety of sources of air pollution, and thus more contributed to the global air pollution. They can emanate from point and nonpoint sources or from mobile and stationary sources. A point source describes a specific location from which large quantities of pollutants are discharged, e.g., coal-fired power plants. A nonpoint source, on the other hand, is more diffuse often involving many small pieces spread across a wide range of area, e.g., automobiles. Automobiles are also known as mobile sources, and the combustion of gasoline is responsible for released emissions from mobile sources. Industrial activities are also known as stationary sources, and the combustion of fossil fuels (coal) is accountable for their emissions.

These pollutants producing from distinct sources may cause harm directly or indirectly. If they are emitted from the source directly into the atmosphere, and so cause harm directly, they are called as primary pollutants, e.g., carbon oxides, carbon monoxide, hydrocarbons, nitrogen oxides, sulfur dioxide, particulate matter, and so on. If they are produced from chemical reactions including also primary pollutants in the atmosphere, they are known as secondary pollutants, e.g., ozone and sulfuric acid.

The Impacts of Air Pollution

The air pollutants result in a wide range of impacts both upon humans and environment. Their

detrimental effects upon humans can be briefly summarized as follows: health problems resulting particularly from toxicological stress, like respiratory diseases such as emphysema and chronic bronchitis, chronic lung diseases, pneumonia, cardiovascular troubles, and cancer, and immune system disorders increasing susceptibility to infection and so on. Their adverse effects upon environment, on the other hand, are the following: acid deposition, climate change resulting from the emission of greenhouse gases, degradation of air resources, deterioration of air quality, noise, photooxidant formation (smog), reduction in the overall productivity of crop plants, stratospheric ozone (O₃) depletion, threats to the survival of biological species, etc.

While determining the extent and degree of harm given by these pollutants, it becomes necessary to know sufficiently about the features of that pollutant. This is because some pollutants can be the reason of environmental or health problems in the air, they can be essential in the soil or water, e.g., nitrogen is harmful as it can form ozone in the air, and it is necessary for the soil as it can also act beneficially as fertilizer in the soil. Additionally, if toxic substances exist below a certain threshold, they are not necessarily harmful.

New Technologies for Air Pollution: Big Data

Before the industrialization period, the components of pollution are thought to be primarily smoke and soot; but with industrialization, they have been expanded to include a broad range of emissions, including toxic chemicals and biological or radioactive materials. Therefore, even today it is still admitted that there are six conventional pollutants (or criteria air pollutants) identified by the US Environmental Protection Agency (EPA): carbon monoxide, lead, nitrous oxides, ozone, particulate matter, and sulfur oxides. Hence, it is expectable that there can be new sources for air pollution and so new threats for the earth soon. Indeed, very recently, through Kigali (Rwanda) Amendment (14 October, 2016) to the Montreal Protocol adopted at the

Meeting of the Parties (MOP 28), it is accepted to address hydrofluorocarbons (HFCs) – greenhouse gases having a very high global warming potential even if not harmful as much as CFCs and HCFCs for the ozone layer under the Protocol – in addition to chlorofluorocarbons (CFCs) and hydrochlorofluorocarbons (HCFCs).

Air pollution first becomes an international issue with the Trail Smelter Arbitration (1941) between Canada and the United States. Indeed, prior to the decision made by the Tribunal, disputes over air pollution between two countries had never been settled through arbitration. Since this arbitration case – specifically with increasing efforts since the early 1990s – attempts to measure, to reduce, and to address rapidly growing impacts of air pollution have been continuing. Developing new technologies, like Big Data, arises as one of those attempts.

Big Data has no uniform definition (ELI 2014; Keeso 2014; Simon 2013; Sowe and Zettsu 2014). In fact, it is defined and understood in diverse ways by different researchers (Boyd 2010; Boyd and Crawford 2012; De Mauro et al. 2016; Gogia 2012; Mayer-Schönberger and Cukier 2013; Manyika et al. 2011) and interested companies like Experian, Forrester, Forte Wares, Gartner, and IBM. It is initially identified by 3Vs – volume (data amount), velocity (data speed), and variety (data types and sources) (Laney 2001). By the time, it has included fourth Vs like veracity (data accuracy) (IBM) and variability (data quality of being subject to structural variation) (Gogia 2012) and a fifth V, value (data capability to turn into value) together with veracity (Marr), and a sixth one, vulnerability (data security-privacy) (Experian 2016). It can be also defined by veracity and value together with visualization (visual representation of data) as additional 3Vs (Sowe and Zettsu 2014) and also by volume, velocity, and variety requiring specific technology and analytical methods for its transformation into value (De Mauro et al. 2016). However, it is generally referred as large and complex data processing sets/applications that conventional systems are not able to cope with them.

Because air pollution has various aspects that should be measured as mentioned above, it

requires massive data that should be collected at different spatial and temporal levels. Therefore, it is observed in practice that Big Data sets and analytics are increasingly used in the field of air pollution, for monitoring, predicting its possible consequences, responding timely to them, controlling and reducing its impacts, and mitigating the pollution itself.

They can be used by different kind of organizations, such as governmental agencies, private firms, and nongovernmental organizations (NGOs). To illustrate, under US Environmental Protection Agency (EPA), samples of Big Data use include:

- Air Quality Monitoring (collaborating with NASA on the DISCOVER-AQ initiative, it involves research on Apps and Sensors for Air Pollution (ASAP), National Ambient Air Quality Standards (NAAQS) compliance, and data fusion methods)
- Village Green Project (on improving Air Quality Monitoring and awareness in communities)
- Environmental Quality Index (EQI) (a dataset consisting of an index of environmental quality based on air, water, land, build environment, and sociodemographic space)

There are also examples generated by local governments like “E-Enterprise for the Environment,” by environmental organizations like “Personal Air Quality Monitoring,” or by citizen science like “Danger Maps,” or by private firms like “Aircraft Emissions Reductions” (ELI 2014) or Green Horizons Project (IBM 2015).

The Environmental Performance Index (EPI) is also another platform – using Big Data compiled from a great number of sensors and models – providing a country and an issue ranking on how each country manages environmental issues and also a Data Explorer allowing users to investigate the global data comparing environmental performance with GDP, population, land area, or other variables.

Despite all, as the potential benefits and costs of the use of Big Data are still under discussion (Boyd 2010; Boyd and Crawford 2012; De Mauro et al. 2016; Forte Wares, –; Keeso 2014; Mayer-Schönberger and Cukier 2013; Simon

2013; Sowe and Zettsu 2014), various concerns can be raised about the use of Big Data to monitor, measure, and forecast air pollution as well. Therefore, it is required to make further research to identify gaps, challenges, and solutions for “making the right data (not just higher volume) available to the right people (not just higher variety) at the right time (not just higher velocity)” (Forte Wares, –).

Cross-References

- Environment
- Pollution, Land
- Pollution, Water

References

- Boyd, Danah. Privacy and publicity in the context of big data. WWW Conference. Raleigh, (2010). Retrieved from <http://www.danah.org/papers/talks/2010/WWW2010.html>. Accession 3 Feb 2017.
- Boyd, Danah & Crawford, Kate. Critical questions for big data, information, communication & society, 15(5), 662–679, (2012). Retrieved from: <http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878>. Accession 3 Feb 2017.
- De Mauro, Andrea, Greco, Marco, Grimaldi, Michele. A formal definition of big data based on its Essential features. (2016). Retrieved from: https://www.researchgate.net/publication/299379163_A_formal_definition_of_Big_Data_based_on_its_essential_features. Accession 3 Feb 2017.
- Environmental Law Institute (ELI). (2014). *Big data and environmental protection: An initial survey of public and private initiatives*. Washington, DC: Environmental Law Institute. Retrieved from: <https://www.eli.org/sites/default/files/eli-pubs/big-data-and-environmental-protection.pdf>. Accession 3 Feb 2017.
- Environmental Performance Index (EPI) (n.d.). Available at: <http://epi.yale.edu/>. Accession 3 Feb 2017.
- Experian. A data powered future. White Paper (2016). Retrieved from: <http://www.experian.co.uk/assets/resources/white-papers/data-powered-future-2016.pdf>. Accession 3 Feb 2017.
- Gartner. Gartner says solving ‘big data’ challenge involves more than just managing volumes of data. June 27, 2011. (2011). Retrieved from: <http://www.gartner.com/newsroom/id/1731916>. Accession 3 Feb 2017.
- Gogia, Sanchit. The big deal about big data for customer engagement, June 1, 2012, (2012). Retrieved from: http://www.iab.fi/media/tutkimus-matskut/130822_

- [forrester_the_big_deal_about_big_data.pdf](#). Accession 3 Feb 2017.
- IBM. IBM expands green horizons initiative globally to address pressing environmental and pollution challenges. (2015). Retrieved from: <http://www-03.ibm.com/press/us/en/pressrelease/48255.wss>. Accession 3 Feb 2017.
- IBM (n.d.). What is big data? Retrieved from: <https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>. Accession 3 Feb 2017.
- Keeso, Alan. Big data and environmental sustainability: A conversation starter. Smith School Working Paper Series, December 2014, Working paper 14-04, (2014). Retrieved from: <http://www.smithschool.ox.ac.uk/library/working-papers/workingpaper%2014-04.pdf>. Accession 3 Feb 2017.
- Laney, D. 3D data management: Controlling data volume, velocity, and variety. Meta Group (2001). Retrieved from: Available at: <https://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Accession 3 Feb 2017.
- Manyika, J. et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute (2011). Retrieved from: https://file:///C:/Users/cassperr/Downloads/MGI_big_data_full_report.pdf. Accession 3 Feb 2017.
- Marr, Bernard (n.d.). Big data: The 5 vs everyone must know. Retrieved from: Available at: <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>. Accession 3 Feb 2017.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how We live, work and think*. London: John Murray.
- Simon, P. (2013). *Too big to ignore: The business case for big data*. Hoboken: Wiley.
- Sowe, S. K. & Zettsu, K. "Curating big data made simple: Perspectives from scientific communities." *Big Data*, 2, 1. 23–33. Mary Ann Liebert, Inc. (2014).
- Wares, F. (n.d.). Failure to launch: From big data to big decisions why velocity, variety and volume is not improving decision making and how to fix it. White Paper. A Forte Consultancy Group Company. Retrieved from http://www.fortewares.com/Administrator/userfiles/Banner/forte-wares-pro-active-reporting_EN.pdf. Accession:3 Feb 2017.
- The Open University. (2007). *T210-environmental control and public health*. Milton Keynes: The Open University.
- Vallero, D. A. (2008). *Fundamentals of air pollution*. Amsterdam: Elsevier.
- Vaughn, J. (2007). *Environmental politics*. Belmont: Thomson Wadsworth.
- Withgott, J., & Brennan, S. (2011). *Environment*. San Francisco: Pearson.

Pollution, Land

Zerrin Savaşan

Department of International Relations,
Sub-Department of International Law, Selçuk
University, Konya, Turkey

Pollution, in its all types (air, water, land), means the entrance of some substances beyond the threshold concentration level into the natural environment which do not naturally belong there and not present there, resulting in its destruction and causing harmful effects on both humans/all living organisms and the environment. So, in land pollution as well, solid or liquid waste materials get deposited on land and further degrade and deteriorate the quality and the productive capacity of land surface. It is sometimes used as a substitute of/or together with soil pollution where the upper layer of the soil is destroyed. However, in fact, soil pollution is just one of the causes of the land pollution.

Like the other types, land pollution also arises as a global environmental problem, specifically associated with urbanization and industrialization, that should be dealt with globally concerted environmental policies. However, as a first and foremost step, it requires to be understood very well with its all dimensions by all humankind, but particularly the researchers studying on it.

What Causes Land Pollution?

The degradation of land surfaces are caused directly or indirectly by human (anthropogenic)

Further Reading

- Gillespie, A. (2006). *Climate change, ozone depletion and air pollution*. Leiden: Martinus Nijhoff Publishers.
- Gurjar, B. R., et al. (Eds.). (2010). *Air pollution, health and environmental impacts*. Boca Raton: CRC Press.
- Jacobson, M. Z. (2012). *Air pollution and global warming*. New York: Cambridge University Press.
- Louka, E. (2006). *International environmental law, fairness, effectiveness, and world order*. New York: Cambridge University Press.
- Raven, P. H., & Berg, L. R. (2006). *Environment*. Danvers: Wiley.

activities. It is possible to mention several reasons temporally or permanently changing the land structure and so causing land pollution. However, three main reasons are generally identified as industrialization, overpopulation, and urbanization, and the others are counted as the reasons stemming from these main reasons. Some of them are as follows: improper waste disposal (agricultural/domestic/industrial/solid/radioactive waste) littering; mining polluting the land through removing the topsoil which forms the fertile layer of soil, or leaving behind waste products and the chemicals used for the process; misuse of land (deforestation, land conversion, desertification); soil pollution (pollution on the topmost layer of the land); soil erosion (loss of the upper (the most fertile) layer of the soil); and the chemicals (pesticides, insecticides, and fertilizers) applied for crop enhancement on the lands.

Regarding these chemicals used for crop enhancement, it should be underlined that, while they are enhancing the crop yield, they can also kill the insects, mosquitoes, and some other small animals. So, they can harm the bigger animals that feed on these tiny animals. In addition, most of these chemicals can remain in the soil or accumulate there for many years. To illustrate, DDT (dichlorodiphenyltrichloroethane) is one of these pesticides. It is now widely banned with the great effect of Rachel Carson's very famous book, *Silent Spring* (1962), which documents detrimental effects of pesticides on the environment, particularly on birds. Nonetheless, as it is not ordinarily biodegradable, so known as persistent organic pollutant, it has remained in the environment ever since it was first used.

Consequences of Land Pollution

All types of pollution are interrelated and their consequences cannot be restricted to the place where the pollution is first discharged. This is particularly because of the atmospheric deposition in which existing pollution in the air (atmosphere) creating pollution in water or land as well.

Since they are interrelated to each other, their impacts are similar to each other as well. Like the

others, land pollution has also serious consequences on both humans, animals and other living organisms, and environment. First of all, all living things depend on the resources of the earth to survive and on the plants growing from the land, so anything that damages or destroys the land ultimately has an impact on the survival of humankind itself and all other living things on the earth. Damages on the land also lead to some problems in relation to health like respiratory problems, skin problems, and various kinds of cancers.

Its effects on environment also require to take attention as it forms one of the most important reasons of the global warming which has started to be a very popular but still not adequately understood phenomena. This emerges from a natural circulation, in turn, land pollution leads to the deforestation, it leads to less rain, eventually to problems such as the greenhouse effect and global warming/climate change. Biomagnification is the other major concern stemming from land pollution. It occurs when certain substances, such as pesticides or heavy metals, gained through eating by aquatic organisms such as fish, which in turn are eaten by large birds, animals, or humans. They become concentrated in internal organs as they move up the food chain, and then the concentration of these toxic compounds tends to increase. This process threatens both these particular species and also all the other species above and below in the food chain. All these combining with the massive extinctions of certain species – primarily because of the disturbance of their habitat – induce also massive reductions in biodiversity.

Control Measures for Land Pollution

Land pollution, along with other types of pollution, poses a threat to the sustainability of world resources. However, while others can have self-purification opportunities through the help of natural events, it can stay as polluted till to be cleaned up. Given the time necessary to pass for the disappearance of plastics in nature (hundreds of years) and the radioactive waste (almost forever), this fact can be understood better. So then land

pollution becomes one of the serious concerns of the humankind.

When the question is asked what should be done to deal with it, first of all, it is essential to remind that it is a global problem having no boundaries, so requires to be handled with collectively. While working collectively, it is first of all necessary to set serious environmental objectives and best-practice measures. A wide range of measures – changing according to the cause of the pollution – can be thought to prevent, reduce, or stop land pollution, such as adopting and encouraging organic farming instead of using chemicals herbicides, and pesticides, restricting or forbidding their usage, developing the effective methods of recycling and reusing of waste materials, constructing proper disposal of all wastes (domestic, industrials, etc.) into secured landfill sites, and creating public awareness and support towards all environmental issues.

Apart from all those measures, the use of Big Data technologies can also be thought as a way of addressing rapidly increasing and wide-ranging consequences of land pollution.

Some of the cases in which Big Data technologies are used in relation to one or more aspects of land pollution can be illustrated as follows (ELI 2014):

- Located under US Department of the Interior (DOI), the National Integrated Land System (NILS) aims to provide the principal data source for land surveys and status by combining Bureau of Land Management (BLM) and Forest Service data into a joint system.
- New York City Open Accessible Space Information System (OASIS) is another sample case; as being an online open space mapping tool, it involves a huge amount of data concerning public lands, parks, community gardens, coastal storm impact areas, and zoning and land use patterns.
- Providing online accession of the state Departments of Natural Resources (DNRs) and other agencies to the data of Geographic Information Systems (GIS) on environmental concerns, while contributing to the effective management of land, water, forest, and wildlife, it

essentially requires the use of Big Data to make this contribution.

- Alabama's State Water Program is another example ensuring geospatial data related to hydrologic, soil, geological, land use, and land cover issues.
- The National Ecological Observatory Network (NEON) is an environmental organization providing the collection of the site-based data related to the effects of climate change, invasive species from 160 sites and also land use throughout the USA.
- The Tropical Ecology Assessment and Monitoring Network (TEAM) is also a global network facilitating the collection and integration of publicly shared data related to patterns of biodiversity, climate, ecosystems, and also land use.
- The Danger Maps is another sample case for the use of Big Data, as it also provides the mapping of government-collected data on over 13,000 polluting facilities in China to allow users to search by area or type of pollution (water, air, radiation, soil).

The US Environmental Protection Agency (EPA) and the Environmental Performance Index (EPI) are also other platforms using Big Data compiled from a great number of sensors regarding environmental issues, on land pollution and on other types of pollution. That is, Big Data technologies can be thought as a way of addressing consequences of all types of pollution, not just of land pollution. This is particularly because, all types of pollution are deeply interconnected with another type, so their consequences cannot be restricted to the place where the pollution is first discharged as mentioned above. Therefore, actually, for all types of pollution, relying on satellite technology and data and data visualization is essentially required to monitor them regularly, to forecast and reduce their possible impacts, and to mitigate the pollution itself. Nonetheless, there are serious concerns raised about different aspects of the use of Big Data in general (boyd 2010; boyd and Crawford 2012; De Mauro et al. 2016; Forte Wares; Keeso 2014; Mayer-Schönberger and Cukier 2013; Simon 2013; Sowe and Zettsu

2014). So, further investigation and analysis are needed to clarify the relevant gaps and challenges regarding the use of Big Data for specifically land pollution.

Cross-References

- [Earth Science](#)
- [Environment](#)
- [Pollution, Air](#)
- [Pollution, Water](#)

Further Reading

- Alloway, B. J. (2001). Soil pollution and land contamination. In R. M. Harrison (Ed.), *Pollution: Causes, effects and control* (pp. 352–377). Cambridge: The Royal Society of Chemistry.
- Boyd, D. (2010). Privacy and publicity in the context of big data. WWW Conference, Raleigh, 29 Apr 2010. Retrieved from <http://www.danah.org/papers/talks/2010/WWW2010.html>. Accessed 3 Feb 2017.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data, information, communication & society. *15*(5), 662–679. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878>. Accessed 3 Feb 2017.
- De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. Retrieved from https://www.researchgate.net/publication/299379163_A_formal_definition_of_Big_Data_based_on_its_essential_features. Accessed 3 Feb 2017.
- Environmental Law Institute (ELI). (2014). *Big data and environmental protection: An initial survey of public and private initiatives*. Washington, DC: Environmental Law Institute. Retrieved from <https://www.eli.org/sites/default/files/eli-pubs/big-data-and-environmental-protection.pdf>. Accessed 3 Feb 2017.
- Environmental Performance Index (EPI). Available at: <http://epi.yale.edu/>. Accessed 3 Feb 2017.
- Forté Wares. Failure to launch: From big data to big decisions why velocity, variety and volume is not improving decision making and how to fix it. White Paper. A Forté Consultancy Group Company. Retrieved from http://www.fortewares.com/Administrator/userfiles/Banner/forte-wares-pro-active-reporting_EN.pdf. Accessed 3 Feb 2017.
- Hill, M. K. (2004). *Understanding environmental pollution*. New York: Cambridge University Press.
- Keeso, A. (2014). Big data and environmental sustainability: A conversation starter. Smith School Working Paper Series, Dec 2014, Working paper 14-04. Retrieved from <http://www.smithschool.ox.ac.uk/library/working-paper/s/workingpaper%2014-04.pdf>. Accessed 3 Feb 2017.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London: John Murray.
- Mirsal, I. A. (2008). *Soil pollution, origin, monitoring & remediation*. Berlin/Heidelberg: Springer.
- Raven, P. H., & Berg, L. R. (2006). *Environment*. Danvers: Wiley.
- Simon, P. (2013). *Too big to ignore: The business case for big data*. Hoboken: Wiley.
- Sowe, S. K., & Zettsu, K. (2014). Curating big data made simple: Perspectives from scientific communities. *Big Data*, 2(1), 23–33. Mary Ann Liebert, Inc.
- Withgott, J., & Brennan, S. (2011). *Environment*. Cornell University: Pearson.

Pollution, Water

Zerrin Savaşan

Department of International Relations,
Sub-Department of International Law, Selçuk
University, Konya, Turkey

Water pollution can be defined as the contamination of water bodies by the entrance of large amounts of materials/substances to those bodies, resulting in physical or chemical change in water, modifying the natural features of the water, degrading the water quality, and adversely affecting the humans and the environment.

Particularly in recent decades, it is highly accepted that water pollution is a global environmental problem which is interrelated to all other environmental challenges. Water pollution control, at national level, generally should involve financial resources, technology improvement, policy measures, and necessary legal and administrative framework and institutional/staff capacity for implementing these policy measures in practice. However, more importantly, at global level, it should involve cooperation of all related sides at all levels. Despite the efforts at both national and global levels, reducing pollution substantially still continues to pose a challenge. This is particularly because even though the world is becoming increasingly globalized, it is still mostly regarded as having with unlimited resources. Hence, it becomes essential to explain that it is

limited and so its resources should not be polluted. Here, it also becomes essential to have adequate information on all types of pollution resulting in environmental deterioration and on water pollution.

What Causes Water Pollution?

This question has many responses, but basically it is possible to mention two main reasons: natural reasons and human-driven reasons. All waters are subject to some degree of natural (or ecological) pollution caused by nature rather than by human activity, through algal blooms, forest fires, floods, sedimentation stemming from rainfalls, volcanic eruptions, and other natural events. However, a greater part of the instances of water pollution arises from humans' activities, particularly from massive industrialization. Accidental spills (e.g., a disaster like the wreck of an oil tanker, as different from others, is unpredictable); domestic discharges; industrial discharges; the usage of large amounts of herbicides, pesticides, chemical fertilizers; sediments in waterways of agricultural fields; improper disposal of hazardous chemicals down the sewages; and being not able to construct adequate waste disposal systems can be expressed as not all but just some of the human-made reasons of water pollution.

The causes as abovementioned vary greatly because a complex variety of pollutants, lying suspended in the water or depositing beneath the earth's surface, get involved in water bodies and result in water quality degradation. Indeed, there are many different types of water pollutants spilling into waterways causing water pollution. They all can be divided up into various categories: chemical, physical, pathogenic pollutants, radioactive substances, organic pollutants, inorganic fertilizers, metals, toxic pollutants, biological pollutants, and so on. Conventional, non-conventional, and toxic pollutants are some of these divisions which are regulated by the US Clean Water Act. The conventional pollutants are as follows: dissolved oxygen, biochemical oxygen demand (BOD), temperature, pH (acid deposition), sewage, pathogenic agents, animal

wastes, bacteria, nutrients, turbidity, sediment, total suspended solids (TSS), fecal coliform, oil, and grease. Nonconventional (or nontoxic) pollutants are not identified as either conventional or priority, like aluminum, ammonia, chloride, colored effluents, exotic species, instream flow, iron, radioactive materials, and total phenols. Toxic pollutants, metals, dioxin, and lead can be counted as examples of priority pollutants. Each group of these pollutants has its own specific ways of entering the water bodies and its own specific risks.

Water Pollution Control

In order to control all these pollutants, it is beneficial to determine from where they are discharged. So, the following categories can be identified to find out where they originate from: point and nonpoint sources of pollution. If the sources causing pollution come from single identifiable points of discharge, they are point sources of pollution, e.g., domestic discharges, ditches, pipes of industrial facilities, and ships discharging toxic substances directly into a water body. Nonpoint sources of pollution are characterized by dispersed, not easily identifiable discharge points, e.g., runoff of pollutants into a waterway, like agricultural runoff, stormwater runoff. As it is harder to identify them, it is nearly impossible to collect, trace, and control them precisely, whereas point sources can be easily controlled.

Water pollution, like other types of pollution, has serious widespread effects. In fact, adverse alteration of water quality produces costs both to humans (e.g., large-scale diseases and deaths) and to environment (e.g., biodiversity reduction, species mortality). Its impact differs depending on the type of water body affected (groundwater, lakes, rivers, streams, and wetlands). However, it can be prevented, lessened, and even eliminated in many different ways. Some of these different treatment methods, aiming to keep the pollutants from damaging the waterways, can be relied on the use of techniques reducing water use, reducing the usage of highly water soluble pesticide and herbicide compounds, and reducing their amounts,

controlling rapid water runoff, physical separation of pollutants from the water, or on the management practices in the field of urban design and sanitation.

There are also some other attempts to measure, reduce, and address rapidly growing impacts of water pollution, such as the use of Big Data. Big Data technologies can provide ways of achieving better solutions for the challenges of water pollution. To illustrate, EPA databases can be accessed and maps can be generated from them including information on environmental activities affecting water and also on air and land in the context of EnviroMapper. Under US Department of the Interior (DOI), National Water Information System (NWIS) monitors surface and underground water quantity, quality, distribution, and movement. Under National Oceanic and Atmospheric Administration (NOAA), California Seafloor Mapping Program (CSMP) works for creating a comprehensive base map series of coastal/marine geology and habitat for all waters of the USA. Additionally, the Hudson River Environmental Conditions Observing System comprises 15 monitoring stations – located between Albany and the New York Harbor – automatically collecting samples every 15 min that are used to monitor water quality, assess flood risk, and assist in pollution cleanup and fisheries management. Contamination Warning System Project, conducted by the Philadelphia Water Department, is a combination of new data technologies with existing management systems. It provides a visual representation of data streams containing geospatial, water quality, customer concern, operations and public health information. Creek Watch is another sample case of the use of Big Data in the field of water pollution. It is developed by IBM and the California State Water Resources Control Board's Clean Water Team as a free app to allow users to rate the waterway on three criteria: amount of water, rate of flow, and amount of trash. The collected data is in large enough to track pollution and manage water resources. The Danger Maps is another project mapping government-collected data on over 13,000 polluting facilities in China. It renders users to search by area or type of pollution (water, air, radiation, soil). Developing

technology on farm performance can also be shown as another sample on the use of Big Data compiled from yield information, sensors, high-resolution maps, and databases for water pollution issue. For example, machine-to-machine (M2M) agricultural technology produced by a Canadian startup company Semios allows farmers to improve yields and their farm operations' efficiency but also it provides information for reducing polluted runoff through increasing the efficient use of water, pesticides, and fertilizers (ELI 2014).

The Environmental Performance Index (EPI) is also another platform using Big Data to display how each country manages environmental issues and to allow users to investigate data through comparing environmental performance with GDP, population, land area, or other variables.

As shown above by example cases, the use of Big Data technologies is increasingly applied in the water field, in its different aspects from management to pollution. However, it is still required to make further research for their effective use in order to eliminate related concerns. This is particularly because there is still debate on the use of Big Data even regarding its general scope and terms (Boyd 2010; Boyd and Crawford 2012; De Mauro et al. 2016; Forte Wares, - ; Keeso 2014; Mayer-Schönberger and Cukier 2013; Simon 2013; Sowe and Zettsu 2014).

Cross-References

- [Earth Science](#)
- [Environment](#)
- [Pollution, Land](#)

Further Reading

- Boyd, D. (2010). Privacy and publicity in the context of big data. WWW Conference, Raleigh, 29 Apr 2010. Retrieved from <http://www.danah.org/papers/talks/2010/WWW2010.html>. Accessed 3 Feb 2017.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data, information, communication & society. *15*(5), 662–679. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878>. Accessed 3 Feb 2017.

- De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. Retrieved from https://www.researchgate.net/publication/299379163_A_formal_definition_of_Big_Data_based_on_its_essential_features. Accessed 3 Feb 2017.
- Environmental Law Institute (ELI). (2014). *Big data and environmental protection: An initial survey of public and private initiatives*. Washington, DC: Environmental Law Institute. Retrieved from <https://www.eli.org/sites/default/files/eli-pubs/big-data-and-environmental-protection.pdf>. Accessed 3 Feb 2017.
- Environmental Performance Index (EPI). Available at: <http://epi.yale.edu/>. Accessed 3 Feb 2017.
- Forte Wares. Failure to launch: From big data to big decisions why velocity, variety and volume is not improving decision making and how to fix it. White Paper. A Forte Consultancy Group Company. Retrieved from http://www.fortewares.com/Administrator/userfiles/Banner/forte-wares-pro-active-reporting_EN.pdf. Accessed 3 Feb 2017.
- Hill, M. K. (2004). *Understanding environmental pollution*. New York: Cambridge University Press.
- Keeso, A. (2014). Big data and environmental sustainability: A conversation starter. Smith School Working Paper Series, Dec 2014, Working paper 14-04. Retrieved from <http://www.smithschool.ox.ac.uk/library/working-papers/workingpaper%2014-04.pdf>. Accessed 3 Feb 2017.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London: John Murray.
- Raven, P. H., & Berg, L. R. (2006). *Environment*. Danvers: Wiley.
- Simon, P. (2013). *Too big to ignore: The business case for big data*. Hoboken: Wiley.
- Sowe, S. K., & Zettsu, K. (2014). Curating big data made simple: Perspectives from scientific communities. *Big Data*, 2(1), 23–33. Mary Ann Liebert, Inc.
- The Open University. (2007). *T210 – Environmental control and public health*. The Open University.
- Vaughn, J. (2007). *Environmental politics*. Thomson Wadsworth.
- Vigil, K. M. (2003). *Clean water: An introduction to water quality and water pollution control*. Oregon State University Press.
- Withgott, J., & Brennan, S. (2011). *Environment*. Pearson.

Precision Agriculture

► [AgInformatics](#)

Precision Farming

► [AgInformatics](#)

Precision Population Health

Emilie Bruzelius^{1,2} and James H. Faghmous¹

¹Arnhold Institute for Global Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

²Department of Epidemiology, Joseph L. Mailman School of Public Health, Columbia University, New York, NY, USA

Synonyms

[Precision public health](#)

Definition

Precision population health refers to the emerging use of big data to improve the health of populations. In contrast to precision medicine, which focuses on detecting and treating disease in *individuals*, precision population health instead focuses on identifying and intervening on the determinants of health within and across *populations*. Though the application of the term “precision” is relatively new, the concept of applying the right intervention at the right time in the right setting is well-established. Recent advances in the volume, accessibility, and ability to process massive datasets offer enhanced opportunities to do just this, holding the potential to monitor population health progress in real-time, and allowing for more agile health programs and policies. As big data and machine learning are increasingly incorporated within population health, future work will continue to focus on the core tasks of improving population-wide prevention strategies, addressing social determinants of health, and reducing health disparities.

What Is Precision Population Health?

Precision population health is a complementary movement to precision medicine that emphasizes the use of big data and other emerging technologies

in advancing population health. These parallel trends in medicine and public health are distinguished by their focus on employing data-driven strategies to predict which health interventions are most likely to be effective for a given individual or population. However, while precision medicine emphasizes clinical applications, often highlighting the explanatory role of genetic differences between individuals, precision population health is oriented towards using data to identify effective interventions for entire populations, often highlighting prevention strategies.

The concept of precision population health is rooted in the precision medicine approach to medical treatment, first successfully pioneered in the context of cancer treatments. Motivated by advances in genetic sequencing, precision medicine tries to take into account an individual's variability in genetics, environment, and behaviors to better optimize therapeutic benefits. The goal of precision medicine is to use data to more accurately predict which treatment is most likely to be effective for a given patient at a given time, rather than treating all patients with the therapy that has been shown to be most effective on average. Central to the notion of precision medicine is the use of large scale data to enhance this process of personalized prediction.

Precision population health, on the other hand, emphasizes the treatment of populations as opposed to individuals, using data-driven approaches to better account for the complex social, environmental, and economic factors that are known to shape patterns of health and illness (Keyes and Galea 2016). In the same way that precision medicine aims to tailor medical treatments to a specific individual's genetics, precision population health aims to tailor public health programs and policies to the needs of specific populations or communities. While not overtly part of the definition, precision population health is understood as contextual – resulting from multiple complex and interacting factors determined not only at the level of the individual, by his or her genetics, health behaviors, and medical care, but also by the broader set of macrosocial forces that accrue over the life-course. In the past several decades, there has been renewed interest in

investigating how these “upstream” factors, shape health distributions, and in using this information to better develop appropriate interventions to prevent disease, promote health, and reduce health disparities. In this context, precision is derived from the use of big data to accomplish population health goals. Enhanced precision is also derived from the use of scalable machine learning algorithms and affordable computing infrastructure to measure exposures, outcomes, and context with granularity and in real time. From this perspective, big data is typically described in the context of the 3 Vs – variety, volume, and velocity – and also includes the broader incorporation of the tools and methods need to manage, store, process and gain insight from such data.

Precision Population Health Opportunities

Recent advances in the volume, variety, and velocity of new data sources provide a unique opportunity to understand and intervene on broad-scale health determinants within and across populations. High volume data refers to the exponential increases, in terms of both rows and columns, of current datasets. This increasing data quantity can provide utility to population health researchers by making new measures available and by increasing sample sizes so that more complex interactions can be evaluated, particularly in terms of rare exposures, outcomes, or subgroups. Further, collecting data at finer spatial and time-scale resolution than has been previously been feasible can help to improve population health program targeting and continuous feedback and program adaptation.

High-variety data refers to the increasing diversity of data types that may be applicable to population health science. These include both traditional sources of epidemiologic data as well as expanding access to newer sources of clinical, administrative, and contextual information. Improved computing power has already facilitated access to rich sources of novel medical data including massive repositories of medical records, imaging, and genetic information.

Other opportunities include administrative sources of information on critical health determinants such as housing, transportation systems, or land-use patterns, that are increasingly available. Remotely sensed data products and weather data may also prove to be of high utility to population health researchers, especially in the context of environmental exposures and infectious disease patterns. Finally, social media content, GPS and other continuous location data, as well as purchasing and transaction data, microfinance and mobile banking information, and wearable technologies offer unique opportunities to study the how social and economic factors shape opportunities and barriers to engaging in health promoting behaviors.

Finally, increasing technological usage is beginning to provide opportunities for high velocity precision population health – close to real-time collection, storage, and analysis of population health data. Instantaneous data collection and analysis, often through the use of algorithms operating without human intervention, holds immense promise for population health monitoring and improvement. These advances may be especially important for population health in the context of continuous monitoring and surveillance activities. For example, the penetration of wearable apps and mobile phone networks over the past decade has expedited the collection health data, also reducing data collection costs by orders of magnitude. These new technologies may prove to be especially important in global settings where national and subnational data on health indicators may not be updates on a routine basis, yet are critical to effective program planning and implementation. Along with mobile data, the use of satellite image analysis is uniquely salient in the context of global population health. For example, recent research has leveraged satellite images of nighttime lights to predict updatable poverty and population density estimates for remote regions (Doupe et al. 2016). Mobile and web technology may also prove to be useful in early detection of anomalies such as disease outbreaks, enabling faster response in times of crisis. More-precise disease surveillance can also generate hypotheses about the causes of emerging disease patterns and

identify early opportunities for prevention. In under-resourced settings, where traditional sources of population data are suboptimal, such methods may provide a useful complement or alternative method for measuring needed population health characteristics for targeted intervention planning and implementation.

Precision Population Health Challenges

The integration of complex population health data poses numerous challenges, many of which, including noise, high dimensionality, and nonlinear relationships, are common in most data-driven explorations (Faghmous 2015). With respect to precision population health however, there are also several unique challenges that should be highlighted. First, though increased access to novel data sources presents new opportunities, working with secondary data can create or reinforce validity challenges as systematic bias due to measurement error cannot be overcome simply with greater volumes of data (Mooney et al. 2015). Though the novel data sources are already beginning to provide insights for global population health programs, these data must be complemented by efforts to expand the collection of population-sampled, representative, health, and demographic data for designing, implementing and monitoring the effectiveness of health based policies and programs. In particular, numerous authors have highlighted the insufficiency of global health data, even with regards to basic metrics such as mortality (Desmond-Hellman 2016).

A second important challenge for precision population health is that a greater emphasis on precision raises potential ethical, social, and legal implications, particularly in terms of privacy. As greater volumes of health data are collected, it will be critical to find ways to protect individual privacy and confidentiality, especially as more data is collected passively through the use of digital services like mobile phones and web searches. Traditional notions of health data privacy, such as those guaranteed under the Health Insurance Portability and Accountability Act (HIPAA), provide data privacy and security provisions for safeguarding

medical information and rely on informed consent for the disclosure and use of an individual's private data. However, regulations regarding the use of nonmedical data are less established, especially with respect to other types of potentially sensitive information and data owned by private sector entities. As discussed, these sources of information may be highly salient to population health researchers. There are serious privacy concerns regarding the use of large-scale patient-level data as the sheer size these datasets increases the risk of potential data breaches by orders of magnitude. At the same time, de-identified datasets may be of limited practical use to clinicians and public health practitioners, especially in the context of health programs that attempt to target high-risk individuals for prevention. The complexity of these issues has led to extensive discussions around the privacy-utility tradeoff of precision population health, yet further work is needed, especially as greater emphasis on scientific collaboration, data sharing, and scientific reproducibility becomes the norm.

Finally, while precision population health holds great promise to improve our ability to predict which health programs and policies are most likely to work, where and for whom, it will be important to continue to focus on core population health tasks, prioritizing population prevention strategies, the role of social and environmental context and addressing health inequity. Much of the current focus on precision has centered too narrowly on genetic and pharmacological factors, rather than on the intersection of precision medicine and precision population health tasks. A better integration of these two themes is critical in order to develop more precise approaches to targeted interventions for both populations and individual patients.

Cross-References

- [Electronic Health Records \(EHR\)](#)
- [Health Informatics](#)
- [Patient-Centered \(Personalized\) Health](#)

References

- Desmond-Hellman, S. (2016). Progress lies in precision. *Science*, 353(6301), 731.
- Doupe, P., Bruzelius, E., Faghmous, J., & Ruchman, S. G. (2016). Equitable development through deep learning: The case of sub-national population density estimation. In *Proceedings of the 7th Annual Symposium on Computing for Development. ACM DEV '16* (pp. 6:1–6:10). New York: ACM.
- Faghmous, J. H. (2015). Machine learning. In A. El-Sayed & S. Galea (Eds.), *Systems science and population health. Epidemiology*. Oxford, UK: Oxford University Press.
- Keyes, K. M., & Galea, S. (2016). Setting the agenda for a new discipline: Population health science. *American Journal of Public Health*, 106(4), 633–634. <https://doi.org/10.2105/AJPH.2016.303101>.
- Mooney, S. J., Westreich, D. J., & El-Sayed, A. M. (2015). Epidemiology in the era of big data. *Epidemiology (Cambridge, Mass.)*, 26(3), 390–394. <https://doi.org/10.1097/EDE.0000000000000274>.

Precision Public Health

- [Precision Population Health](#)
-

Predictive Analytics

Anamaria Berea

Department of Computational and Data Sciences,
George Mason University, Fairfax, VA, USA

Center for Complexity in Business, University of
Maryland, College Park, MD, USA

Predictive analytics is a methodology in data mining that uses a set of computational and statistical techniques to extract information from data with the purpose to predict trends and behavior patterns. Often, the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown data, whether it is in the past, present, or future (Siegel 2013). In other words, predictive analytics can be applied not only to time series data but to any data where there is some unknown that can be inferred.

Therefore prediction analytics is a powerful set of tools for inferring lost past data as well.

The core of predictive analytics in data science relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions (Tukey 1977).

Predictive Analytics and Forecasting

Prediction, in general, is about forecasting the future or forecasting the unknown. In the past, before the scientific method was invented, predictions were based on astrological observations, witchcraft, foretelling, oral history folklore, and, in general, on random observations or associations of observations that happened at the same time. For example, if a conflict happened during an eclipse, then all eclipses would become “omens” of wars and, in general, bad things. For a long period of time in our civilization, the events were merely separated in two classes: good or bad. And thus the associations of events that would lead to a major conflict or epidemics or natural catastrophe would be categorized as “bad” omens from there on, while any associations of events that would lead to peace, prosperity, and, in general, “good” major events would be categorized as “good” omens or good predictors from there on.

The idea of associations of events as predictive for another event is actually at the core of some of the statistical methods we are using today, such as correlation. But the fallacy of using these methods metaphorically instead of in a quantitative systematic analysis is that only one set of observations cannot be predictive for the future. That was true in the past and it is true now as well, no matter how sophisticated the techniques we are using. Predictive analytics uses a series of events or associations of events, and the longer the series, the more informative the predictive analysis can be.

Unlike past good or bad omens, the results of predictive analytics are probabilistic. This means that predictive analytics informs the probability of a certain data point or the probability of a hypothesis to be true.

While true prediction can be achieved only by determining clearly the cause and the effect in a set of data, a task that is usually hard to do, most of the predictive analytics techniques are outputting probabilistic values and error term analyses.

Predictive Modeling Methods

Predictive modeling statistically shows the underlying relationships in historical, time series data in order to explain the data and make predictions, forecasts, or classifications about future events.

In general, predictive analytics uses a series of statistical and computational techniques in order to forecast future outcomes from past data. Traditionally, the most used method has been the linear regression, but lately, with the emergence of the Big Data phenomenon, there have been developed many other techniques aiming to support businesses and forecasters, such as machine learning algorithms or probabilistic methods.

Some classes of techniques include:

1. Applications of both linear and nonlinear mathematical programming algorithms, in which one objective is optimized within a set of constraints.
2. Advanced “neural” systems, which learn complex patterns from large datasets to predict the probability that a new individual will exhibit certain behaviors of business interest. Neural networks (also known as deep learning) are biologically inspired machine learning models that are being used to achieve the recent record-breaking performance on speech recognition and visual object recognition.
3. Statistical techniques for analysis and pattern detection within large datasets.

Some techniques in predictive analytics are borrowed from traditional forecasting techniques,

such as moving average, linear regressions, logistic regressions, probit regressions, multinomial regressions, time series models, or random forest techniques. Other techniques, such as supervised learning, A/B testing, correlation ranking, k-nearest neighbor algorithm are closer to machine learning and newer computational methods.

One of the most used techniques in predictive analytics today though is supervised learning or supervised segmentation (Provost and Fawcett 2013). Supervised segmentation includes the following steps:

- Selection of informative attributes – particularly in large datasets, the selection of the variables that are more likely to be informative to the goal of prediction is crucial; otherwise the prediction can render spurious results.
- Information gain and entropy reduction – these two techniques measure the information in the selected attributes.
- Selection is done based on tree induction, which fundamentally represents subsetting the data and searching for these informative attributes.
- The resulting tree-structured model partitions the space of all data into possible segments with different predicted values.

The supervised learning/segmentation has been popular because it is computationally and algorithmically simple.

Visual Predictive Analytics

Data visualization and predictive analytics complement each other nicely and together they are an even more powerful methodology for the analysis and forecasting of complex datasets that comprise a variety of data types and data formats.

Visual predictive analytics is a specific set of techniques of predictive analytics that is applied to visual and image data. Just as in the case of predictive analytics in general, temporal data is

required for the visual (spatial) data (Maciejewski et al. 2011). This technique is particularly useful in determining hotspots and areas of conflict with a high dynamics. Some of the techniques used in spatiotemporal analysis are kernel density estimation for event distribution and seasonal trend decomposition by loess smoothing (Maciejewski et al. 2011).

Predictive Analytics Example

A good example for using predictive analytics is in healthcare. The problem of understanding the probability of an upcoming epidemics or the probability of increase in incidence of various diseases, from flu to heart disease and cancer.

For example, given a dataset that contains data with respect to the past incidence of heart disease in the USA, demographic data (gender, average income, age, etc.), exercise habits, eating habits, traveling habits, and other variables, a predictive model would follow these steps:

1. Descriptive statistics – the first step in doing predictive analytics or building a predictive model is always an understanding of the data with respect to what the variables represent, what ranges they fall into, how long is the time series, ASO, essentially a summary statistics of the data.
2. Data cleaning and treatment – it is very important to understand not what the data is or has but also what the data is missing.
3. Build the model/s – in this step, several techniques can be explored and used comparatively and based on their results; the best one should be chosen. For example, both a general regression and a random forest can be used and compared, or supervised segmentation based on demographics and then the segments compared.
4. Performance and accuracy estimation – in this final step, the probabilities or measurements of forecasting accuracy are computed and interpreted.

In any predictive model or analytics technique, the model can do only what the data is. In other words, it is impossible to assess a predictive model of the heart disease incidence based on the travel habits if no data regarding travel is included.

Another important point to remember is that the accuracy of the model also depends on the accuracy measure, and using multiple accuracy measures is desired (i.e., mean squared error, p-value, R-squared).

In general, any predictive analytic technique will output a dataset of created variables, called predictive values, and the newly created dataset. Therefore a good technique for verification and validation of the methods used is to partition the real dataset in two sets and use one to “train” the model and the second one to validate the model’s results.

The success of the model ultimately depends on how real events will unfold and that is one of the reasons why longer time series are better at informing predictive modeling and giving better accuracy for the same set of techniques.

Predictive Analytics Fallacies

Cases of “spurious correlations” tend to be quite famous, such as the correlation between the number of people who dies tangled in their bed sheets and the consumption of cheese per capita (<http://www.tylervigen.com/spurious-correlations>). These examples fall on the same fallacy as the “bad”/“good” omen one, as the observations of the events at the same time does not imply that there is a causal relationship between the two events.

Another classic example is to think, in general, that correlations show a causal relationship; therefore predictions based on correlation analyses alone tend to fail often.

Some other fallacies of predictive analytics techniques include an insufficient analysis of the errors, relying on the p-value alone, relying on a Poisson distribution of the current data, and many more.

Predictive/Descriptive/Prescriptive

There is a clear distinction between descriptive vs. predictive vs. prescriptive analytics in Big Data (Shmueli 2010). Descriptive analytics shows how past or current data can be analyzed in order to determine patterns and extract meaningful observations out of the data. Predictive analytics is generally based on a model that is informed by descriptive analytics and gives various outcomes based on past data and the model. Prescriptive analytics is closely related to predictive analytics, as it takes the predictive values, puts them in a decision model, and informs the decision-makers about the future course of action (Shmueli and Koppius 2010).

Predictive Analytics Applications

In practice, predictive analytics can be applied to almost all disciplines – from predicting the failure of mechanical engines in hard sciences, to predicting customers’ buying power in social sciences and business (Gandomi and Haider 2015).

Predictive analytics is especially used in business and marketing forecasting. Hair Jr. (2007) shows the importance of predictive analytics for marketing and how it has become more relevant with the emergence of the Big Data phenomenon. He argues that survival in a knowledge-based economy is derived from the ability to convert information to knowledge. Data mining identifies and confirms relationships between explanatory and criterion variables. Predictive analytics uses confirmed relationships between variables to predict future outcomes. The predictions are most often values suggesting the likelihood a particular behavior or event will take place in the future.

Hair also argues that, in the future, we can expect predictive analytics to increasingly be applied to databases in all fields and revolutionize the ability to identify, understand, and predict future developments; data analysts will increasingly rely on mixed-data models that examine

both structured (numbers) and unstructured (text and images) data; statistical tools will be more powerful and easier to use; future applications will be global and real time; demand for data analysts will increase as will the need for students to learn data analysis methods; and scholarly researchers will need to improve their quantitative skills so the large amounts of information available can be used to create knowledge instead of information overload.

Predictive Modeling and Other Forecasting Techniques

Some predictive modeling techniques do not necessarily involve Big Data. For example, Bayesian networks and Bayesian inference methods, while they can be informed by Big Data, they cannot be applied granularly to each data point due to the computational complexity that can arise from calculating thousands of conditional probability tables. But Bayesian models and inferences can certainly be used in combination with statistical predictive modeling techniques in order to bring the analysis closer to a cause-effect type of inference (Pearl 2009).

Another forecasting technique, that does not rely on Big Data, but harnesses the power of the crowds, is the prediction market. Just like Bayesian modeling, prediction markets can be used as a complement to Big Data and predictive modeling in order to augment the likelihood value of the predictions (Arrow et al. 2008).

Cross-References

► [Business Intelligence Analytics](#)

References

- Arrow, K.J., et al. (2008). *The promise of prediction markets*. Science-New York then Washington-320.5878: 877.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.

- Hair Jr., J. F. (2007). Knowledge creation in marketing: The role of predictive analytics. *European Business Review*, 19(4), 303–315.
- Maciejewski, R., et al. (2011). Forecasting hotspots – A predictive analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 17(4), 440–453.
- Pearl, J. (2009). *Causality*. Cambridge: Cambridge university press.
- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol: O'Reilly Media.
- Shmueli, G. (2010) To Explain or to Predict?. *Statistical Science* 25(3):289–310.
- Shmueli, G., & Koppius, O. (2010). Predictive analytics in information systems research. Robert H. Smith School Research Paper No. RHS, 06-138.
- Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Hoboken: Wiley.
- Tukey, J. (1977). *Exploratory data analysis*. New York: Addison-Wesley.

Prevention

David Brown^{1,2} and Stephen W. Brown³

¹Southern New Hampshire University, University of Central Florida College of Medicine, Huntington Beach, CA, USA

²University of Wyoming, Laramie, WY, USA

³Alliant International University, San Diego, CA, USA

One of the primary purposes of government is to provide for the health and safety of its constituents. From humanitarian, economic, and public health perspectives, prevention is the most effective and efficient approach towards achieving this goal. Research and program evaluation studies repeatedly demonstrate that prevention activities improve health and safety outcomes. Health and safety problem prevention are much safer, efficient, and cost-effective than health and safety problem treatment.

Since its development, the discipline of public health has had disease, accident, illness, and harm prevention as one of its primary goals. Program

S. W. Brown: deceased.

evaluation studies have demonstrated that effective prevention efforts improve health outcomes, and they lower the cost of health care for both program participants and nonparticipants.

All aspects of the public health prevention model have been advanced through the use of big data. Big data greatly enhances the ability to identify environmental, genetic, and lifestyle factors that might increase or decrease the risk of diseases, illness, and accidents. Big data increases the speed at which new vaccines and other prevention programs can be developed and evaluated. Big data dramatically improves the ability to identify large geographic areas and highly specific locations at risk of illnesses, accidents, crimes, and epidemics.

In public health activities, the adage is more information makes for better programs. Thus, the belief of public health scientists that more information has been generated in the last 5 years than in the entire history of mankind leads to the conclusion that big data has the potential of leading to great strides in the advancement of public health prevention programs.

Big data is enhancing epidemiologists' abilities to identify risk factors that increase the probability of diseases, illnesses, accidents, and other types of harm. This information is being used to develop programs designed to decrease or eliminate the identified risks. Big data helps clinical researchers track the efficacy of their treatments; this knowledge is used to develop new interventions designed to prevent other clinical problems. Accident and crime prevention experts use big data to predict areas likely to suffer accidents and/or criminal behavior. This information is being used to lower crime rates and improve accident statistics. Big data mechanisms can be used to track and map the spread of infectious diseases. This information has significant implications for worldwide disease prevention and health improvement.

Electronic medical records and online patient charting systems are frequently used sources of prevention big data. As an example, anonymous aggregate data from these systems help identify gaps, disparities, and unnecessary duplications in healthcare delivery. This information has a cost-

saving function, and the data can be used to evaluate people's responses to prevention programs and activities.

Global Positioning System (GPS) big data information is being used to bring emergency care to areas in need of first responder services. This system has led to significant decreases in emergency response time. Fast response time often prolongs life and prevents further complications from emergency situations. Communities that have installed GPS systems have also seen a significant decrease in accidents involving their first responders and their vehicles.

Big data is also being used to help people who suffer from chronic diseases. As an example, in the case of asthma, a wireless sensor is attached to the patient's medication inhaler. This sensor provides information about the amount of medication being administered, the time of administration, and the location of the patient using the inhaler. A smartphone is then used to transmit the information to care providers and researchers. This program has led to significant decreases in the incidence of uncontrolled asthma attacks.

Big data is being used by community police departments to track the time and place of incidents of accidents and crimes. The use of such data has led to significant decreases in accidents and the incidence of violent crime in many communities.

Big data facilitates the elimination of risk factors that contribute to the development of chronic disease such as diabetes, obesity, and heart disease. Wearable monitors can assess physical activity, diet, tobacco use, drug use, and exposure to pollution. These data then lead to the discovery and prevention of risk factors for public health problems at the population, subpopulation, and individual levels. It can improve peoples' quality of life by monitoring intervention effectiveness and by helping people live healthier lives in healthier environments.

Conclusion

As technology continues to advance, additional opportunities will present themselves to utilize

big data techniques to prevent disease and disability around the world. As additional big data sources and technologies develop, it is reasonable to predict a decrease in their cost and increase their effectiveness. However, as in all systems, while the data itself is highly valuable, it is not the data that is the primary source of improved prevention activities and programs. Rather, it is information gleaned from the data and the questions that the data answer is of most value. The effective use of big data has great potential to prevent illness, accidents, diseases, and crimes that cause harm to the public good on a worldwide scale. Big data and big improvements in disease, accident, illness, and harm prevention would definitely seem to go hand in hand.

Cross-References

- [Biomedical Data](#)
- [Electronic Health Records \(EHR\)](#)
- [Evidence-Based Medicine](#)
- [Health Care Delivery](#)
- [Participatory Health and Big Data](#)
- [Patient-Centered \(Personalized\) Health](#)

Further Reading

- Barrett, M. Humblet, O., Hiatt, R.A., et al. (2013). Big data and disease prevention. *Big Data* September 2013.
- Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28 (Suppl 3), 660–665.
- Hay, S. I., George, D. B., Moyes, C. L., & Brownstein, J. S. (2013). Big data opportunities for global infectious disease surveillance. *PLoS Medicine*, 10(4), 1–4. <https://doi.org/10.1371/journal.pmed.1001413>.
- Michael, K., & Miller, K. W. (2013). Big data: New opportunities and new challenges. *Computer*, 46(6), 22–24.
- Van Sickle, D., Maenner, M., Barrett, M., et al. (2013). Monitoring and improving compliance and asthma control: Mapping inhaler use for feedback to patients, physicians and payers. *Respiratory Drug Delivery Europe*, 1, 1–12.

Privacy

Joanna Kulesza

Department of International Law and
International Relations, University of Lodz, Lodz,
Poland

Origins and Definition

Privacy is a universally recognized human right, subject to state protection from arbitrary or unlawful interference and unlawful attacks. The age of Big Data has brought it to the foreground of all technology-related debates as the amount of information aggregated online, generated by various sources together with the computing capabilities of modern networks, makes it easy to connect an individual to a particular piece of information about them, possibly causing a direct threat to their privacy. Yet international law grants every person the right to legal safeguards against any interference with one's right or attacks upon it. The right to privacy covers, although is not limited to, one's identity, integrity, intimacy, autonomy, communication, and sexuality and results in legal protection for one's physical integrity; health information, including sex orientation and gender; reputation; image; personal development; personal autonomy; and self-determination as well as family, home, and correspondence that are to be protected by state from arbitrary or unlawful interferences by its organs or third parties. This catalogue is meant to remain an open one, enabling protection of forever new categories of data, such as geographical location data or arguably to a "virtual personality." As such, the term covers also information about an individual that is produced, generated, or needed for the purpose of rendering electronic services, such as a telephone, an IMEI or an IP number, e-mail address, a website address, geolocation data, or search terms, as long as such information may be linked to an individual and allows for their identification. Privacy is not an absolute right and may be limited for reasons considered necessary in a

democratic society. While there is no *numerus clausus* of such limitative grounds, they usually include reasons of state security and public order or the rights of others, such as their freedom of expression. States are free to introduce certain limitations on individual privacy right as long as those are introduced by specific provisions of law, communicated to the individuals whose privacy is impacted, and applied solely when necessary in particular circumstances. This seemingly clear and precise concept suffers practical limitations as states differ in their interpretations of “necessity” of interference as well as the “specificity” of legal norms required and scope of their application. As a consequence the concept of privacy strongly varies throughout the world’s regions and countries. This is a particular challenge at the time of Big Data as various national and regional perceptions of privacy need to be applied to the very same vast catalogue of online information.

This inconsistency in privacy perceptions results from varied cultural and historical background of individual states as well as their differing political and economic situation. In countries recognizing values reflected in universal human rights treaties, including Europe, large parts of Americas, and some Asian states, the right to privacy covers numerous elements of individual autonomy and is strongly protected by comprehensive legal safeguards. On the other hand in countries rapidly developing, as well as in ones with unstable political or economic situation, primarily located in Asia and Africa, the significance of the right to one’s private life subsides to urgent needs of protecting life and personal or public security. As a consequence the undisputed right to privacy, subject to numerous international treaties and rich international law jurisprudence, remains highly ambiguous, an object of conflicting interpretations by national authorities and their agents. This is one of the key challenges to finding the appropriate legal norms governing Big Data. In the unique Big Data environment, it is not only the traditional jurisdictional challenges, specific to all online interactions, that must be faced but also the tremendously varying perceptions of privacy

all finding their application to the vast and varied Big Data resource.

History

The idea of privacy rose simultaneously in various cultures. Contemporary authors most often refer to the works of American and European legal writers of late nineteenth century to identify its origins. In US doctrine it were Warren and Brandeis who introduced in their writings “the right to be let alone,” a notion still often used to describe the essential content of privacy. Yet at remotely the same time, German legal scholar Kohler published a paper covering a similar concept. It was also in mid nineteenth century that French courts issued their first decisions protecting the right to private life. The right to privacy was introduced to grant individuals protection from undesired intrusions into their private affairs and home life, be it by nosy journalists or governmental agents. Initially the right was used to limit the rapidly evolving press industry, with time, as individual awareness and recognition of the right increased; the right to privacy primarily introduced limits of individual information that state or local authorities may obtain and process. As any new idea, the right to privacy initially provoked much skepticism, yet by mid twentieth century became a necessary element of the rising human rights law. In the twenty-first century, it gained increased attention as a side effect of the growing, global information society. International online communications allowed for easy and cheap mass collection of data, creating the greatest threat to privacy so far. What followed was an eager debate on the limits of allowed privacy intrusions and actions required from states aimed at safeguarding the rights of an individual. A satisfactory compromise is not easy to find as states and communities view privacy differently, based on their history, culture, and mentality. The existing consensus on human rights seems to be the only starting point of a successful search for an effective privacy compromise, much needed in the era of transnational companies operating on

Big Data. With the modern notions of “the right to be forgotten” or “data portability” referring to new facets of the right to protect one’s privacy, the Big Data phenomenon is one of the deciding factors of this ongoing evolution.

Privacy as a Human Right

The first document of international human rights law, recognizing the right to privacy, was the 1948 Universal Declaration on Human Rights (UDHR). The nonbinding political middle ground was not too difficult to find with the greatest horrors in human history of World War II still vividly in the mind of world’s politicians and citizens alike. With horrid memories fading away and the Iron Curtain drawing a clear line between differing values and interests, a binding treaty on the very issue took almost 20 more years. Irreconcilable differences between communist and capitalist countries covered the scope and implementation of individual property, free speech, or privacy. The eventual 1966 compromise in the form of the two fundamental human rights treaties: the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic Social and Cultural Rights (ICESCR) allowed for a conciliatory wording on hard law obligations for different categories of human rights, yet left the crucial details to future state practice and international jurisprudence. Among the right to be put into detail by future state practice, international courts, and organizations was the right to privacy, established as a human right in Article 12 UDHR and Article 17 ICCPR. They both granted every individual freedom from “arbitrary interference” with their “privacy, family, home, or correspondence” as well as from any attacks upon their honor and reputation. While neither document defines “privacy,” the UN Human Rights Committee (HRC) has gone into much detail on delimitating its scope for the international community. All 168 ICCPR state parties are obliged per the Covenant to reflect HRC recommendations on the scope and enforcement of the treaty in general and privacy in particular. Over time the HRC produced detailed

instruction on the scope of privacy protected by international law, discussing the thin line with state sovereignty, security, and surveillance. According to Article 12 UDHR and Article 17 ICCPR, privacy must be protected against “arbitrary or unlawful” intrusions or attacks through national laws and their enforcement. Those laws are to detail limits for any justified privacy invasions. Those limits of individual privacy right are generally described in Article 29 para. 2 which allows for limitations of all human rights determined by law solely for the purpose of securing due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality, public order, and the general welfare in a democratic society. Although proposals for including a similar restraint in the text of the ICCPR were rejected by negotiating parties, the right to privacy is not an absolute one. Following HRC guidelines and state practice surrounding the ICCPR, privacy may be restrained according to national laws which meet the general standards present in human rights law. The HRC confirmed this interpretation in its 1988 General Comment No. 16 as well as recommendations and observations issued thereafter. Before Big Data became, among its other functions, an effective tool for mass surveillance, the HRC took a clear stand on the question of legally permissible limits of state inspection. It clearly stated that any surveillance, whether electronic or otherwise; interceptions of telephonic, telegraphic, and other forms of communication; wiretapping; and recording of conversations should be prohibited. It confirmed that individual limitation upon privacy must be assessed on a case-by-case basis and follow a detailed legal guideline, containing precise circumstances when privacy may be restricted by actions of local authorities or third parties. The HRC specified that even interference provided for by law should be in accordance with the provisions, aims, and objectives of the Covenant and reasonable in the particular circumstances, where “reasonable” means justified by those particular circumstances. Moreover, as per the HRC interpretation, states must take effective measures to guarantee that information about individual’s life

does not reach ones not authorized by law to obtain, store, or process it. Those general guidelines are to be considered the international standard of protecting the human right to privacy and need to be respected regardless of the ease that Big Data services offer in connecting pieces of information available online with individuals they relate to. Governments must ensure that Big Data is not to be used in a way that infringes individual privacy, regardless of the economic benefits and technical accessibility of Big Data services.

The provisions of Article 17 ICCPR resulted in similar stipulations of other international treaties. Those include Article 8 of the European Convention on Human Rights (ECHR) binding upon its 48 member states or Article 11 of the American Convention on Human Rights (ACHR) agreed upon by 23 parties to the treaty. The African Charter on Human and Peoples' Rights (Banjul Charter) does not contain a specific stipulation regarding privacy, yet its provisions of Article 4 on the inviolability of human rights, Article 5 on human dignity, and Article 16 on the right to health serve as basis to grant individuals within the jurisdiction of 53 state parties the protection recognized by European or American states as inherent to the right of privacy. While no general human rights document exists among Australasian states, the general guidelines provided by the HRC and the work of the OECD are often reflected in national laws on privacy, personal rights, and personal data protection.

Privacy and Personal Data

The notion of personal data is closely related to that of privacy, yet their scopes differ. While personal data is a term relatively well defined, privacy is a more broad and ambiguous notion. As Kuner rightfully notes, the concept of privacy protection is a broader one than personal data regulations, where the latter provides a more detailed framework for individual claims. The influential Organization for Economic Co-operation and Development (OECD) Forum identified personal data as a component of the

individual right to privacy, yet its 34 members differ on the effective methods of privacy protection and the extent to which such protection should be granted. Nevertheless, the nonbinding yet influential 1980 OECD Guidelines on the Protection of Privacy and Transborder Flow of Personal Data (Guidelines) together with their 2013 update have so far encouraged over data protection laws in over 100 countries, justifying the claim that, thanks to its detailed yet unified character and national enforceability personal data protection, is the most common and effective legal instrument safeguarding individual privacy. The Guidelines identify the universal privacy protection through eight personal data processing principles. The definition of "personal data" contained in the Guidelines is usually directly adopted by national legislations which cover any information relating to an identified or identifiable individual, referred to as "data subject." The basic eight principles of privacy and data protection include (1) the collection limitation principle, (2) the data quality principle, (3) the individual participation principle, (4) the purpose specification principle, (5) the use limitation principle, (6) the security safeguards principle, (7) the openness principle, and (8) the accountability principle. They introduce certain obligations upon "data controllers" that is parties "who, according to domestic law, are competent to decide about the contents and use of personal data regardless of whether or not such data are collected, stored, processed or disseminated by that party or by an agent on their behalf." They oblige data controllers to respect limits made by national laws pertaining to the collection of personal data. As already noted this is of particular importance to Big Data operators, who must be aware and abide by the varying national regimes. Personal data must be obtained by "lawful and fair" means and with the knowledge or consent of the data subject, unless otherwise provided by relevant law. Collecting or processing personal data may only be done when it is relevant to the purposes for which it will be used. Data must be accurate, complete, and up to date. The purposes for data collection ought to be specified no later than at the time of data collection. The use of the data must be

limited to the purposes so identified. Data controllers, including those operating on Big Data, are not to disclose personal data at their disposal for purposes other than those initially specified and agreed upon by the data subject, unless such use or disclosure is permitted by law. All data processors are to show due diligence in protecting their collected data, by introducing reasonable security safeguards against the loss or unauthorized data access and its destruction, use, modification, or disclosure. This last obligation may prove particularly challenging for Big Data operators, with regard to the multiple locations of data storage and their continuously changeability. Consequently each data subjects enjoys the right to obtain information on the fact of the data controller having data relating to him, to have any such data communicated within a reasonable time, to be given reasons if a request for such information is denied, as well as to be able to challenge such denial and any data relating to him. Followingly each data subject enjoys the right to have their data erased, rectified, completed, or amended, and data controller is to be held accountable to national laws for lack of effective measures ensuring all of those personal data rights.

Therewith the OECD principles form a practical standard for privacy protection represented in the human rights catalogue, applicable also to Big Data operators, given the data in their disposal relates directly or indirectly to an individual. While their effectiveness may come to depend upon jurisdictional issues, the criteria for identification of data subjects and obligations of data processors are clear.

Privacy as a Personal Right

Privacy is recognized not only by international law treaties and international organizations but also by national laws, from constitutions to civil and criminal law codes and acts. Those regulations hold great practical significance, as they allow for direct remedies against privacy infractions from private parties, rather than those

enacted by state authorities. Usually privacy is considered an element of the larger catalogue of personal rights and granted equal protection. It allows individuals whose privacy is under threat for the threatening activity to be seized (e.g., infringing information be deleted or a press release be stopped). It also allows for pecuniary compensation or damages should a privacy infringement already take place.

Originating from German-language civil law doctrine, privacy protection may be well described by the theory of concentric spheres. Those include the public, private, and intimate sphere, with different degrees of protection from interference granted to each of them. The strongest protection is granted to intimate information; activities falling within the public sphere are not protected by law and may be freely collected and used. All individual information may be qualified as falling into one of the three spheres, with the activities performed in the public sphere being those performed by an individual as a part of their public or professional duties and obligations and deprived of privacy protection. This sphere would differ as per individual, with “public figures” enjoying least protection. An assessment of the limits of one’s privacy when compared with their public function would always be made on case-by-case basis. Any information that may not be considered public is to be granted privacy protection and may only be collected or processed with permission granted by the one it concerns. The need to obtain consent from the individual the information concerns is also required for the intimate sphere, where the protection is even stronger. Some authors argue that information on one’s health, religious beliefs, sexual orientation, or history should only be distributed in pursuit of a legitimate aim, even when permission for its distribution was granted by the one it concerns.

With the civil law scheme for privacy protection being relatively simple, its practical application relies on case-by-case basis and therefore may show challenging and unpredictable in practice, especially when international court practice is of issue.

Privacy and Big Data

Big Data is a term that directly refers to information about individuals. It may be defined as gathering, compiling, and using large amounts of information enabling for marketing or policy decisions. With large amounts of data being collected by international service providers, in particular ones offering telecommunication services, such as Internet access, the scope of data they may collect and the use to which they may put it is of crucial concern to all their clients but also to their competitors and state authorities interested in participating in this valuable resource. In the light of the analysis presented above, any information falling within the scope of Big Data that is collected and processed while rendering online services may be considered subject to privacy protection when it refers to identified or identifiable individual that is a physical person who may either be directly identified or whose identification is possible. When determining whether particular category or a piece of information constitutes private data, account must be taken of means likely reasonably to be used either by any person to identify the individual, in particular costs, time, and labor needed to identify such person. When private information has been identified, the procedures required for privacy protection described above ought to be applied by entities dealing with such information. In particular the guidelines described by the HRC in their comments and observations may serve as a guideline for handling personal data falling within the Big Data resource. Initiatives such as Global Network Initiative, a bottom-up initiative of the biggest online service providers aimed at identifying and applying universal human rights standards for online services, or the UN Protect Respect and Remedy Framework for business, defining the human rights obligations of private parties, present a useful tool for introducing enhanced privacy safeguards for all Big Data resources. With the users' growing awareness of the value of their privacy, company privacy policies prove to be a significant element of the marketing game,

inciting Big Data operators to convince forever more users to choose their privacy-oriented services.

Summary

Privacy recognized as a human right requires certain precautions to be taken by state authorities and private business alike. Any information that may allow for the identification of an individual ought to be subjected to particular safeguards allowing for its collection or processing solely based on the consent of the individual in question or a particular norm of law applicable in a case where the inherent privacy invasion is reasonable and necessary to achieve a justifiable aim. In no case may private information be collected or processed in bulk, with no judicial supervision or without the consent of the individual it refers to. Big Data offer new possibilities for collecting and processing personal data. When designing Big Data services or using information they provide, all business entities must address the international standards of privacy protection, as identified by international organizations and good business practice.

Cross-References

- ▶ [Data Processing](#)
- ▶ [Data Profiling](#)
- ▶ [Data Provenance](#)
- ▶ [Data Quality Management](#)
- ▶ [Data Security](#)
- ▶ [Data Security Management](#)

Further Reading

- Kuner, C. (2009). An international legal framework for data protection: Issues and prospects. *Computer Law and Security Review*, 25(263), 307.
- Kuner, C. (2013). *Transborder data flows and data privacy law*. Oxford: Oxford University Press.
- UN Human Rights Committee. General Comment No. 16: Article 17 (Right to Privacy), The Right to Respect of

Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation. 8 Apr 1988. <http://www.refworld.org/docid/453883f922.html>.

UN Human Rights Council. Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, Martin Scheinin. U.N. Doc. A/HRC/13/37.

Warren, S.D., & Brandeis, L.D. (1980). The right to privacy. *Harvard Law Review*, v. 4/193.

Weber, R.H. (2013). Transborder data transfers: Concepts, regulatory approaches and new legislative initiatives. *International Data Privacy Law* v. 1/3–4.

Probabilistic Matching

Ting Zhang

Department of Accounting, Finance and Economics, Merrick School of Business, University of Baltimore, Baltimore, MD, USA

Definition/Introduction

Probabilistic matching differs from the simplest data matching technique, deterministic matching. For deterministic matching, two records are said to match if one or more identifiers are identical. Deterministic record linkage is a good option when the entities in the data sets have identified common identifiers with a relatively high quality of data. Probabilistic matching is a statistical approach in measuring the probability that two records represent the same subject or individual based on whether they agree or disagree on the various identifiers (Dusetzina et al. 2014).

It calculates linkage composite weights based on likeness scores for identifier values and uses thresholds to determine a match, nonmatch, or possible match. The quality of resulting matches can depend upon one's confidence in the specification of the matching rules (Zhang and Stevens 2012). It is designed to work using a wider set of data elements and all available identifiers for matching and does not require identical identifier or exact matches. Instead, it compares the probability of a match to a chosen threshold.

Why Probabilistic Matching?

Although deterministic matching is important in the big data world, it works well with high quality data. However, often data we have has no known or identical identifiers with missing, incomplete, erroneous, or inaccurate values. Some data may change over time, such as address changes due to relocation or name changes due to marriage or divorce. Sometimes there could be a typos, words out of order, split words, extraneous or missing, or wrong information in identification number (see Zhang and Stevens 2012).

In the big data world, larger data sets have more attributes involved and more complex rules-based matching routines. In that case, implementation deterministic matching can involve many man hours of processing, testing, customization, and revision time and longer deployment times than probabilistic matching. As Schumacher (2007) mentioned, unlike probabilistic matching that has scalability and capability to perform lookups in real time, deterministic matching does not have speed advantages.

As Schumacher (2007) suggested, probabilistic matching assign a probability of the quality of a match allowing variation and nuances, it is better suited for complex data systems with multiple databases. Larger databases often mean greater potential for duplicates, human error, and discrepancies; this makes the matching technique designed to determine links between records with complex error patterns more effective. For probabilistic matching, users decide the tolerance level of their choice for a match.

Steps for Probabilistic Matching

This matching technique typically includes three stages: pre-matching data cleaning, matching stage, post-matching data manual review. For the match stage, Dusetzina et al. (2014) summarize the probabilistic matching steps as follows:

1. Estimate the *match* and *non-match* probabilities for each linking variable using the

observed frequency of agreement and disagreement patterns among all pairs, commonly generated using the expectation-maximization algorithm described by Fellegi and Sunter (1969). The *match probability* is the probability of agreed identifier, and the *non-match probability* is the probability that false matches randomly agree on the identifier.

2. Calculate agreement and disagreement weights using the *match* and *non-match* probabilities. The weight assigned to agreement or disagreement on each identifier is assessed as a likelihood ratio, comparing the *match probability* to the *non-match probability*.
3. Calculate a total linking weight for each pair by summing the individual linking weights for each linkage variable.
4. Compare the total linkage weight to a chosen threshold above which pairs are considered a link. The threshold is set using information generated in Step 1.

Applications

Data Management

Probabilistic matching is used to create and manage databases. It helps to clean, reconcile data, and remove duplicates.

Data Warehousing and Business Intelligence

Probabilistic matching plays a key role in data warehousing. This method can help merge multiple datasets from various sources into one.

Medical History and Practice

Medical data warehouse put together using probabilistic matching can help quickly extract a patient's medical history for better medical practice.

Longitudinal Study

Data warehouse based on probabilistic matching can be used to put together longitudinal datasets for longitudinal studies.

Software

Link Plus

One often used free software is *Link Plus*, developed by the Centers for Disease Control and Prevention. *Link Plus* is a probabilistic record linkage software product originally designed to be used by cancer registries. However, *Link Plus* can be used with any type of data and has been used extensively across diverse research disciplines.

The Link King

The Link King is another free software, but it requires a license for base SAS. It is developed by Washington State's Division of Alcohol and Substance Abuse. Like *Link Plus*, the software provides a straightforward user interface using information including first and last names.

Other Public Software

ChoiceMaker and *Freely Extensible Biomedical Record Linkage (FEBRL)* are two publicly available software that health services researchers have used frequently in recent years (Dusetzina et al. 2014). *Record Linkage At IStat (RELAIS)* is a JAVA, R, and MySQL based open source software.

Known Commercial Software

Selected commercial softwares include *LinkageWiz*, *G-Link* developed by Statistics Canada based on Winkler (1999), *LinkSolv*, *Strategic-Matching*, and *IBM InfoSphere Master Data Management* for enterprise data.

Conclusion

Probabilistic matching is a statistical approach in measuring the probability that two records represent the same subject or individual based on whether they agree or disagree on the various identifiers. It has superiority over simplistic deterministic matching. The method itself follows several steps. Its application includes data management, data warehousing, medical practice, and longitudinal research. A variety of public and

commercial software to conduct probabilistic matching is available.

Further Readings

- Dusetzina, S. B., Tyree, S., Meyer, A. M., et al. (2014). *Linking data for health services research: A framework and instructional guide*. Rockville: Agency for Healthcare Research and Quality (US).
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183–1210.
- Schumacher, S. (2007). *Probabilistic versus deterministic data matching: Making an accurate decision, information management special reports*. Washington, DC: The Office of the National Coordinator for Health Information Technology (ONC).
- Winkler, W. E. (1999). *The state of record linkage and current research problems*. Washington, DC: Statistical Research Division, US Census Bureau.
- Zhang, T., & Stevens, D. W. (2012). Integrated data system person identification: Accuracy requirements and methods. <https://ssrn.com/abstract=2512590>; <https://doi.org/10.2139/ssrn.2512590>.

Profiling

Patrick Juola

Department of Mathematics and Computer Science, McAnulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, PA, USA

Profiling is the analysis of data to determine features of the data source that are not explicitly present in the data. For example, by examining information related to a specific criminal act, investigators may be able to determine the psychology and the background of the perpetrator. Similarly, advertisers may look at public behavior to identify psychological traits, with an eye to targeting ads to more effectively influence individual consumer's behavior. This has proven to be a controversial application of big data both for ethical reasons and because the effectiveness of profiling techniques has been questioned.

Profiling is sometimes distinguished from identification (see De-identification/Re-identification) because what is produced is not a specific

individual identity, but a set of characteristics that can apply to many people, but is still useful. One application is in criminal investigations. Investigators use profiling to identify characteristics of offenders based on what is known of their actions (Douglas and Burgess 1986). For example, the use of specific words by anonymous letter writers can help link different letters to the same person and in some cases can provide deeper information. In one case (Shuy 2001), an analysis of a ransom note turned up an unusual phrase indicating that the writer of the note was from the Akron, Ohio area; this knowledge made it easy to identify the actual kidnapper from among the suspects.

Unfortunately, this kind of specific clue is not always present at the crime scene and may require specialist knowledge to interpret. Big data provides one method to fill this gap by treating profiling as a data classification/machine learning problem and analyzing large data sets to learn differences among classes, then applying this to specific data of interest.

For example, the existence of gender differences in language is well-known (Coates 2015). By collecting large samples of writing by both women and men, a computer can be trained to learn these differences and then determine the gender of the unknown author of a new work (Argamon et al. 2009). Similar analyses can determine gender, age, native language, and even personality traits (Argamon et al. 2009). Other types of analysis, such as looking at Facebook “likes,” can evaluate a person's traits more accurately than the person's close friends (Andrews 2018).

This kind of knowledge can be used in many ways beyond law enforcement. Advertisements, for example, can be more effective when tailored to the recipient's traits (Andrews 2018). However, this lends itself to data abuses, such as Cambridge Analytica's attempt to manipulate elections, including the 2016 US Presidential election and the 2016 UK Brexit referendum. Using personality-based microtargeting, the company suggested different advertisements to persuade individual voters to vote in the desired way (Rathi 2019).

This has been described as an “ethical grey area” and an “[attempt] to manipulate voters by latching onto their vulnerabilities” (Rathi 2019). However, it is also not clear whether or not the

models used were accurate enough to be effective, or how many voters were actually persuaded to cast their votes in the correct way (Rathi 2019).

As with any active research area, the performance and effectiveness of profiling are likely to progress over time. Debates on the ethics, effectiveness, and even legality of this sort of profile-based microtargeting are likely to continue for the foreseeable future.

Cross-References

► [De-identification/Re-identification.](#)

Further Reading

- Andrews, E. L. (2018) *The science behind Cambridge analytica: Does psychological profiling work? Insights by Stanford Business.* <https://www.gsb.stanford.edu/insights/science-behind-cambridge-analytica-does-psychological-profiling-work>.
- Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119–123.
- Coates, J. (2015). *Women, men and language: A sociolinguistic account of gender differences in language*. New York: Routledge.
- Douglas, J. E., & Burgess, A. E. (1986). Criminal profiling: A viable investigative tool against violent crime. *FBI Law Enforcement Bulletin*, 55(12), 9–13. <https://www.ncjrs.gov/pdffiles1/Digitization/103722-103724NCJRS.pdf>.
- Rathi, R. (2019). Effect of Cambridge analytica's Facebook ads on the 2016 US Presidential election. *Towards Data Science.* <https://towardsdatascience.com/effect-of-cambridge-analyticas-facebook-ads-on-the-2016-us-presidential-election-dacb5462155d>.
- Shuy, R. W. (2001). DARE's role in linguistic profiling, 4 DARE Newsletter 1.

Psychology

Daniel N. Cassenti and Katherine R. Gamble
U.S. Army Research Laboratory, Adelphi,
MD, USA

Wikipedia introduces big data as “a blanket term for any collection of data sets so large and

complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.” The field of psychology is interested in big data in two ways: (1) at the level of the data, that is, how much data there are to be processed and understood, and (2) at the level of the user, or how the researcher analyzes and interprets the data. Thus, psychology can serve the role of helping to improve how researchers analyze big data and provide data sets that can be examined or analyzed using big data principles and tools.

Psychology

Psychology may be divided into two overarching areas: clinical psychology with a focus on individuals, and the fields of experimental psychology with foci on the more general characteristics that apply to the majority of people. Allen Newell classifies the fields of experimental psychology by time scale, to include biological at the smallest time scale, cognitive (the study of mental processes) at the scale of hundreds of milliseconds to tens of seconds, rational (the study of decision making and problem solving) at minutes to hours, and social at days to months. The cognitive, rational, and social bands can all be related to big data in terms of both the researcher analyzing data and the data itself. Here, we describe how psychological principles can be applied to the researcher to handle data in the cognitive and rational fields and demonstrate how psychological data in the social field can be big data.

Cognitive and Rational Fields

One of the greatest challenges of big data is its analysis. The principles of cognitive and rational psychology can be applied to improve how the big data researcher evaluates and makes decisions about the data. The first step in analysis is attention to the data, which often involves filtering out irrelevant from relevant data. While many software programs can provide an automated filtering of data, the researcher must still give attention and critical analysis to the data as a check on the

automated system, which operates within rigid criteria preset by the researcher that is not sensitive to the context of the data. At this early level of analysis, the researcher's perception of the data, ability to attend and retain attention, and working memory capacity (i.e., the quantity of information that an individual can store while working on a task) are all important to success. That is, the researcher must efficiently process and highlight the most important information, stay attentive enough to do this for a long period of time, and because of limited working memory capacity and a lot of data to be processed, effectively manage the data, such as by chunking information, so that it is easier to filter and store in memory.

The goal of analysis is to lead to decisions or conclusions about data, the scope of the rational field. If all principles from cognitive psychology have been applied correctly (e.g., only the most relevant data are presented and only the most useful information stored in memory), tenets of rational psychology must next be applied to make good decisions about the data. Decision making may be aided by programming the analysis software to present decision options to the researcher. For example, in examining educational outcomes of children who come from low income families, the researcher's options may be to include children who are or are not part of a state-sponsored program, or are of a certain race. Statistical software could be designed to present these options to the researcher, which may reveal results or relationships in the data that the researcher may not have otherwise discovered. Option presentation may not be enough, however, as researchers must also be aware of the consequences of their decisions. One possible solution is the implementation of associate systems for big data software. An associate system is automation that attempts to advise the user, in this case to aid decision making. Because these systems are knowledge based, they have situational awareness and are able to recommend courses of action and the reasoning behind those recommendations. Associate systems do not make decisions themselves, but instead work semiautonomously, with the user imposing supervisory control. If the researcher

deems recommended options to be unsuitable, then the associate system can present what it judges to be the next best options.

Social Field

The field of social psychology provides good examples of methods of analysis that can be used with big data, especially with big data sets that include groups of individuals and their relationships with one another, the scope of social psychology. The field of social psychology is able to ask questions and collect large amounts of data that can be examined and understood using these big data-type analyses including, but not limited to, the following types of analyses.

Linguistic analysis offers the ability to process transcripts of communications between individuals, or to groups as in social media applications, such as tweets from a Twitter data set. A linguistic analysis may be applied in a multitude of ways, including analyzing the qualities of relationship between individuals or how communications to groups may differ based on the group. These analyses can determine qualities of these communications, which may include trust, attribution of personal characteristics, or dependencies, among other considerations.

Sentiment analysis is a type of linguistic analysis that takes communications and produces ratings of the emotional valence individuals direct to the topic. This is of value for considerations of social data researchers who must find those with whom alliances may be formed and who to avoid. A famous example is the strategy shift taken by United States Armed Forces commanders to ally with Iraqi residents. Sentiment analysis indicated which residential leaders would give their cooperation for short-term goals of mutual interest.

The final social psychological big data analysis technique under consideration here is social-network analysis or SNA. With SNA, special emphasis is not with the words spoken as in linguistic and sentiment analysis but on the directionality and frequency of communication

between individuals. SNA created a type of network map that uses nodes and ties to connect members of groups or organizations to one another. This visualization tool allows a researcher to see how individuals are connected to one another with factors like the thickness of a line to determine frequency of communication, or the number of lines coming from a node determining the number of nodes to which they are connected.

Psychological Data as Big Data

Each field of psychology potentially includes big data sets for analysis by a psychological researcher. Traditionally, psychologists have collected data on a smaller scale using controlled methods and manipulations analyzable with traditional statistical analyses. However, with the advent of big data principles and analysis techniques, psychologists can expand the scope of data collection to examine larger data sets that may lead to new and interesting discoveries. The following section discusses each of the aforementioned fields.

In clinical psychology, big data may be used to diagnose an individual. In understanding an individual or attempting to make a diagnosis, the person's writings and interview transcripts may be analyzed in order to provide insight to his or her state of mind. To thoroughly analyze and treat a person, a clinical psychologist's most valuable tool may be this type of big data set.

Biological psychology includes the subfields of psychophysiology and neuropsychology. Psychophysiological data may include hormone collection (typically salivary), blood flow, heart rate, skin conductance, and other physiological responses. Neuropsychology includes multiple technologies for collecting information about the brain, including electroencephalography (EEG), functional magnetic resonance imaging (fMRI), functional near infrared spectroscopy (fNIRS), among other lesser used technologies. Measures in biological psychology are generally taken near-

continuously across a certain time range, so much of the data collected in this field could be considered big data.

Cognitive psychology covers all mental processing. That is, this field includes the initiation of mental processing from internal or external stimuli (e.g., seeing a stoplight turn yellow), the actual processing of this information (e.g., understanding that a yellow light means to slow down), and the initiation of an action (e.g., knowing that you must step on the brake in order to slow your car). For each action that we take, and even actions that may be involuntary (e.g., turning your head toward an approaching police siren as you begin to slow your car), cognitive processing must take place at the levels of perception, information processing, and initiation of action. Therefore, any behavior or thought process that is measured in cognitive psychology will yield a large amount of data for even the simplest of these, such that complex processes or behaviors measured for their cognitive process will yield data sets of the magnitude of big data.

Another clear case of a field with big data sets is rational psychology. In rational psychological paradigms, researchers who limit experimental participants to a predefined set of options often find themselves limiting their studies to the point of not capturing naturalistic rational processing. The rational psychologist, instead typically confronts big data as imaginative solutions to problems, and many forms of data, such as verbal protocols (i.e., transcripts of participants explaining their reasoning), require big data analysis techniques.

Finally, with the large time band under consideration, social psychologists must often consider days' worth of data in their studies. One popular technique is to have participants use wearable technology to periodically remind them to record how they are doing, thinking, and feeling during the day. These types of studies lead to big data sets not just because of the frequency with which the data is collected, but also due to the enormous number of possible activities, thoughts, and feeling that participants may have experienced and recorded at each prompted time point.

The Unique Role of Psychology in Big Data

As described above, big data plays a large role in the field of psychology, and psychology can play an important role in how big data are analyzed and used. One aspect of this relationship is the necessity of the role of the psychology researcher on both ends of big data. That is, psychology is a theory-driven field, where data are collected in light of a set of hypotheses, and analyzed as either supporting or rejecting those hypotheses. Big data offers endless opportunities for exploration and discovery in other fields, such as creating word clouds from various forms of social media to determine what topics are trending, but solid psychological experiments are driven by a priori ideas, rather than data exploration. Thus, psychology is important to help big data researchers learn how to best process their data, and many types of psychological data can be big data, but the importance of theory, hypotheses, and the role of the researcher will always be integral in how psychology and big data interact.

Cross-References

- [Artificial Intelligence](#)
- [Communications](#)
- [Decision Theory](#)

- [Social Media](#)
- [Social Network Analysis](#)
- [Social Sciences](#)
- [Socio-spatial Analytics](#)
- [Visualization](#)

Further Reading

- Cowan, N. (2004). *Working memory capacity*. New York: Psychology Press.
- Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. In *Situation awareness analysis and measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis*. Cambridge, MA: MIT-press.
- Lewis, T. G. (2011). *Network science: Theory and applications*. Hoboken: Wiley.
- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco: W.H. Freeman and Co.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–35.
- Pentland, A. (2014). *Social physics: How good ideas spread – The lessons from a new science*. New York: Penguin Press.
- Yarkoni, T. (2012). Psychoinformatics new horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science*, 21(6), 391–397.

Recommender Systems

Julian McAuley
Computer Science Department, UCSD, San
Diego, USA

Introduction

Every day we interact with predictive systems that seek to model our behavior, monitor our activities, and make recommendations: Whom will we befriend? What articles will we like? What products will we purchase? Who influences us in our social network? And do our activities change over time? Models that answer such questions drive important real-world systems, and at the same time are of basic scientific interest to economists, linguists, and social scientists, among others.

Recommender Systems aim to solve tasks such as those above, by learning from large volumes of historical activities to describe the dynamics of user preferences and the properties of the content users interact with. Recommender systems can take many forms (Table 1), though in essence all boil down to modeling the *interactions between users and content*, in order to predict future actions and preferences. In this chapter, we investigate a few of the most common models and paradigms, starting with item-to-item recommendation (e.g., “people who like x also like y ”), followed by systems that model user preferences and item properties, and finally systems that make

use of rich content, such as temporal information, text, or social networks.

Scalability issues are a major consideration when applying recommender systems in industrial or other “big data” settings. The systems we describe below are those specifically designed to address such concerns, through use of sparse data structures and efficient approximation schemes, and have been successfully applied to real-world applications including recommendation on *Netflix* (Bennett and Lanning 2007), *Amazon* (Linden et al. 2003), etc.

Preliminaries & Notation. We consider the scenario where users (U) interact with items (I), where “interactions” might describe purchases, clicks, likes, etc. (In certain instances (like friend recommendation) the “user” and “item” sets may be the same.) In this setting, we can describe users’ interactions with items in terms of a (sparse) matrix:

$$A = \left(\begin{array}{ccccc} & \overbrace{1 & 0 & \cdots & 1}^{\text{items}} \\ \begin{array}{c} 1 \\ 0 \\ \vdots \\ 1 \end{array} & \begin{array}{c} 0 \\ 0 \\ \ddots \\ 0 \end{array} & \begin{array}{c} \cdots \\ \ddots \\ \cdots \end{array} & \begin{array}{c} 1 \\ 0 \\ \vdots \\ 1 \end{array} \end{array} \right) \Bigg\} \text{users}, \quad (1)$$

where $A_{ui} = 1$ if and only if the user u interacted with the item i . A row of the matrix A_u is now a binary vector describing which items the user u interacted with, and a column $A_{\cdot,i}$ is a binary vector describing which users interacted with the

Recommender Systems, Table 1 Different types of recommender system (for a hypothetical fashion recommendation scenario). (U = user; I = item; F = feature space; I^* = sequence of items)

Output type	Example/applications	Example input/output
$f: U \times I \rightarrow I$	Item-to-Item recommendation & collaborative filtering	user $\times$  $\rightarrow$ 
$f: U \times I \rightarrow \mathbb{R}$	Model-based recommendation (rating prediction)	user $\times$  $\rightarrow$ 
$f: U \times I \times F \rightarrow \mathbb{R}$	Content/context-aware recommendation	user $\times$  $\rightarrow$  [color:red, size:12, price:\$80] [gender:f, location:billings-MT]
$f: U \times I^* \times I \rightarrow \mathbb{R}$	Temporal/sequence-aware recommendation	user $\times$    $\rightarrow$  1/5/15 4/7/15 6/11/15 recent_purchases

item i . Equivalently, we can describe interactions in terms of sets:

$$I_u = \{i | A_{u,i} = 1\} \quad (2)$$

$$U_i = \{u | A_{u,i} = 1\}. \quad (3)$$

Such data are referred to as *implicit feedback*, in the sense that we observe only what items users interacted with, rather than their preferences toward those items. In many cases, interactions may be associated with *explicit feedback* signals, e.g., numerical scores such as star ratings, which we can again describe using a matrix:

$$R = \begin{pmatrix} 4 & ? & \dots & 3 \\ ? & ? & & ? \\ \vdots & & \ddots & \vdots \\ 2 & ? & \dots & 1 \end{pmatrix}. \quad (4)$$

Note that the above matrix is *partially observed*, that is, we only observe ratings for those items the users interacted with. We can now describe recommender systems in terms of the above matrices, e.g., by estimating interactions A_{ui} that are likely to occur, or by predicting ratings R_{ui} .

Models

Item-to-Item Recommendation and Collaborative Filtering. Identifying relationships among items is a fundamental part of many real-world recommender systems, e.g., to generate

recommendations of the form “people who like x also like y .” To do so, a system must identify which items i and j are *similar* to each other.

In the simplest case, “similarity” might be measured by counting the overlap between the set of users who interacted with the two items, e.g., via the *Jaccard Similarity*:

$$\text{Jaccard}(i, j) = \frac{|U_i \cap U_j|}{|U_i \cup U_j|}. \quad (5)$$

Note that this measure takes a value between 0 (if no users interacted with both items) and 1 (if exactly the same set of users interacted with both items). Where explicit feedback is available, we might instead measure the similarity between users’ rating scores, e.g., via the *Pearson Correlation*:

$$\begin{aligned} \text{Cor}(i, j) &= \frac{\sum_{u \in U_i \cap U_j} (R_{u,i} - \bar{R}_{\cdot,i})(R_{u,j} - \bar{R}_{\cdot,j})}{\sqrt{\sum_{u \in U_i \cap U_j} (R_{u,i} - \bar{R}_{\cdot,i})^2 \sum_{u \in U_i \cap U_j} (R_{u,j} - \bar{R}_{\cdot,j})^2}}, \end{aligned} \quad (6)$$

which takes a value from 1 (both items were rated by the same set of users, and those users had the same opinion polarity about them), and -1 (both items were rated by the same users, but users had the *opposite* opinion polarity about them).

Simple similarity measures such as those above can be used to make recommendations by identifying the items j (from some candidate set)

that are most similar to the item i currently being considered:

$$\operatorname{argmax}_j \operatorname{Cor}(i, j) \quad (7)$$

‘Model-Based’ Recommendation. Model-based recommender systems attempt to estimate user “preferences” and item “properties” so as to directly optimize some objective, such as the error incurred when predicting the rating $r(u, i)$ when the true rating is $R_{u,i}$, e.g., via the *Mean Squared Error* (MSE):

$$\frac{1}{|R|} \sum_{u,i \in R} (r(u, i) - R_{u,i})^2. \quad (8)$$

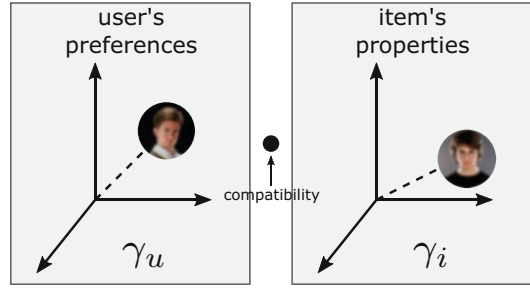
A trivial form of model-based recommender might simply associate each item with a bias term β_i (how good is the item?) and each user with a bias term β_u (how generous is the user with their ratings?), so that ratings would be predicted by

$$r(u, i) = \alpha + \beta_u + \beta_i, \quad (9)$$

where α is a global offset. A more complex system might capture interactions between a user and an item via multidimensional user and item terms:

$$r(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i, \quad (10)$$

where γ_u and γ_i are low-rank matrices that describe *interactions* between the user u and the item i in terms of the user’s preferences and the item’s properties. This idea is depicted in Fig. 1. The dimensions or “factors” that describe an item’s properties (γ_i) might include (for example) whether a movie has good special effects, and the corresponding user factor (γ_u) would capture whether the user cares about special effects; their inner product then describes whether the user’s preferences are “compatible” with the item’s properties (and will thus give the movie a high rating). However, no “labels” are assigned to the factors; rather the dimensions are discovered simply by factorizing the matrix R in terms of the low-rank factors γ_i and γ_u . Thus, such models are



Recommender Systems, Fig. 1 Latent-factor models describe users’ preferences and items’ properties in terms of low-dimensional factors

described as *latent factor models* or *factorization-based approaches*.

Finally, the parameters must be optimized so as to minimize the MSE:

$$\begin{aligned} & \alpha, \beta, \gamma \\ & = \operatorname{argmin}_{\alpha, \beta, \gamma} \frac{1}{|R|} \sum_{u,i \in R} (\alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i - R_{u,i})^2. \end{aligned} \quad (11)$$

This can be achieved via gradient descent, i.e., by computing the partial derivatives of Eq. (11) with respect to α , β , and γ , and updating the parameters iteratively.

Variants and Extensions

Temporal Dynamics and Sequential Recommendation. Several works extend recommendation models to make use of timestamps associated with feedback. For example, early similarity-based methods (e.g., Ding and Li 2005) used time-weighting schemes that assign decaying weights to previously rated items when computing similarities. More recent efforts are frequently based on matrix factorization, where the goal is to model and understand the historical *evolution* of users and items, via temporally evolving offsets, biases, and latent factors (e.g., parameters $\beta_u(t)$ and $\gamma_u(t)$ become functions of the timestamp t). For example, the winning solution to the *Netflix* prize (Bennett and Lanning 2007) was largely based on a series of insights that extended matrix

factorization approaches to be temporally aware (Koren et al. 2009). Variants of *temporal recommenders* have been proposed that account for short-term bursts and long-term “drift,” user evolution, etc.

Similarly, the order or *sequence* of activities that users perform can provide informative signals, for example, knowing what action was performed most recently provides context that can be used to predict the next action. This type of “first-order” relationship can be captured via a Markov Relationship, which can be combined with factorization-based approaches (Rendle et al. 2010).

One-Class Collaborative Filtering. In many practical situations, explicit feedback (like ratings) are not observed, and instead only implicit feedback instances (like clicks, purchases, etc.) are available. Simply training factorization-based approaches on an implicit feedback matrix (A) proves ineffective, as doing so treats “missing” instances as being inherently negative, whereas these may simply be items that a user is unaware of, rather than items they explicitly dislike. The concept of *One-Class Collaborative Filtering* (OCCF) was introduced to deal with this scenario (Pan et al. 2008). Several variants exist though a popular approach consists of sampling pairs of items i and i' for each user u (where i was clicked/purchased and i' was not) and maximizing an objective of the form

$$\underbrace{u, i, i'}_{\text{sample}} \sum \ln \sigma(r(u, i) - r(u, i')). \quad (12)$$

Optimizing such an objective encourages items i (with which the user is likely to interact) to have a larger scores compared to items i' (that they are unlikely to interact with) (Rendle et al. 2009).

Content-Aware Recommendation. So far, the systems we have considered only make use of interaction data, but ignore *features* associated with the users and items being considered. *Content-aware recommenders* can improve the performance of traditional approaches, especially in “cold-start” situations where few interactions are associated with users and items.

For example, suppose we are given binary features associated with a user (or equivalently an item), $A(u)$. Then, we might fit a model of the form

$$r(u, i) = \alpha + \beta_u + \beta_i + \left(\gamma_u + \sum_{a \in A(u)} \rho_a \right) \cdot \gamma_i \quad (13)$$

where ρ_a is a vector of parameters associated with the a^{th} attribute (Koren et al. 2009). Essentially, ρ_a in this setting determines how the our estimate of the user’s preference vector (γ_u) changes as a result of having observed the attribute a (which might correspond to a feature like age or location). Variants of such models exist that make use of rich and varied notions of “content,” ranging from locations to text and images.

Cross-References

► Collaborative Filtering

References

- Bennett, J., & Lanning, S. (2007). The Netflix prize. In *KDD Cup and Workshop*.
- Ding, Y., & Li, X. (2005). Time weight collaborative filtering. In *CIKM*. ACM. <https://dl.acm.org/citation.cfm?id=1099689>.
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30–37. <https://dl.acm.org/citation.cfm?id=1608614>.
- Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1), 76–80. <https://ieeexplore.ieee.org/document/1167344/>.
- Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R., Scholz, M., & Yang, Q. (2008). One-class collaborative filtering. In *ICDM*. IEEE. <https://dl.acm.org/citation.cfm?id=1511402>.
- Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. In *UAI*. AUAI Press. <https://dl.acm.org/citation.cfm?id=1795167>.
- Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized Markov chains for next-basket recommendation. In *WWW*. ACM. <https://dl.acm.org/citation.cfm?id=1772773>.

Regression

Qinghua Yang

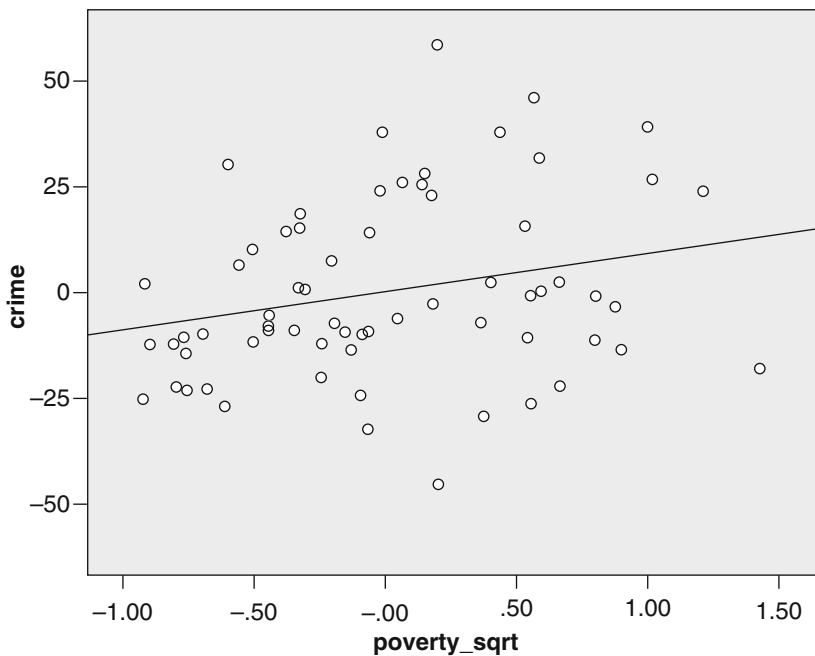
Department of Communication Studies, Texas
Christian University, Fort Worth, TX, USA

Regression is a statistical tool to estimate the relationship(s) between a dependent variable (y or outcome variable) and one or more independent variables (x or predicting variables; Fox 2008). More specifically, regression analysis helps in understanding the variation in a dependent variable using the variation in independent variables with other confounding variable(s) controlled. Regression analysis is widely used to make prediction and estimation of the conditional expectation of the dependent variable given the independent variables, where its use overlaps with the field of machine learning. Figure 1 shows how crime rate is related to residents' poverty level and predicts the crime rate of a specific community.

We know from this regression that there is a positive linear relationship between the crime rate (y axis) and residents' poverty level (x axis). Given the poverty index of a specific community, we are able to make a prediction of the crime rate at that area.

Linear Regression

The estimation target of regression is a function that predicts the dependent variable based upon values of the independent variables, which is called the regression function. For **simple linear regressions**, the function can be represented as $y_i = \alpha + \beta x_i + \varepsilon_i$. The function of **multiple linear regressions** is $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i$ where k is the number of independent variables. The regression estimation using **ordinary least squares (OLS)** selects the line with the lowest total sum of squared residuals. The proportion of total variation (SST) that is explained by the regression (SSR) is known as the **coefficient**



Regression, Fig. 1 Linear regression of crime rate and residents' poverty level

of determination, often referred to as R^2 , a value ranging between 0 and 1 with a higher value indicating a better regression model (Keith 2015).

Nonlinear Regression

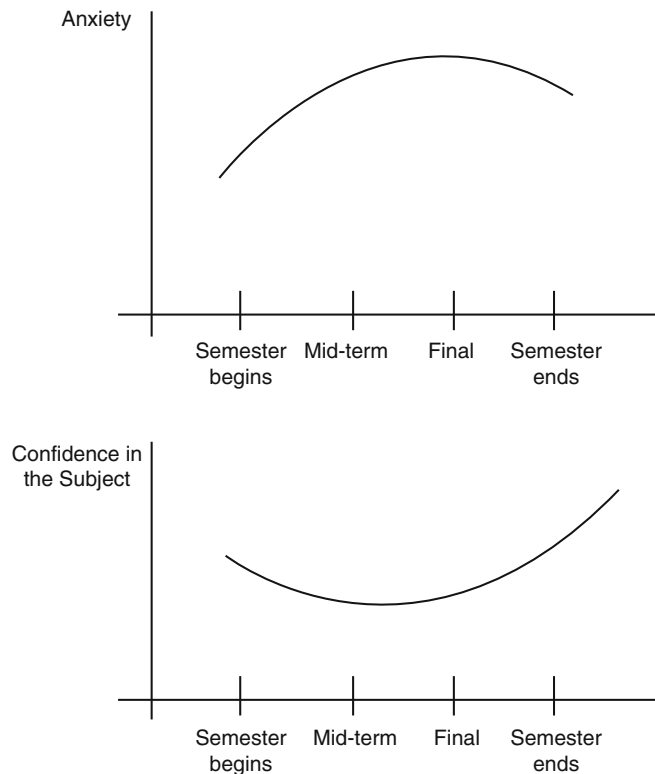
In the real world, there are much more nonlinear functions than linear ones. For example, the relationship between x and y can be fitted in a **quadratic function** shown in Figure 2. There are in general two ways to deal with nonlinear models. First, nonlinear models can be approximated with linear functions. Both nonlinear functions in Figure 2 can be approximated by two linear functions according to the slope: the first linear regression function is from the beginning of the semester to the final exam, and the second function is from the final to the end of the semester. Similarly, regarding cubic, quartic, and more complicated regressions, they can also be approximated with a sequence of linear functions.

However, analyzing nonlinear models in this way can produce much residual and leave considerable variance unexplained. The second way is considered better than the first one from this aspect, by including nonlinear terms in the regression function as $\hat{y} = \alpha + \beta_1x + \beta_2x^2$. As the graph of a quadratic function is a parabola, if $\beta_2 < 0$, the parabola opens downward, and if $\beta_2 > 0$, the parabola opens upward. Instead of having x^2 in the model, the nonlinearity can also be presented in many other ways, such as $\sqrt{x}$, $\ln(x)$, $\sin(x)$, $\cos(x)$, and so on. However, which nonlinear model to choose should be based on both theory or former research and the R^2 .

Logistic Regression

When the outcome variable is dichotomous (e.g., yes/no, success/failure, survived/died, accept/reject), **logistic regression** is applied to make prediction of the outcome variable. In logistic

Regression,
Fig. 2 Nonlinear
regression models



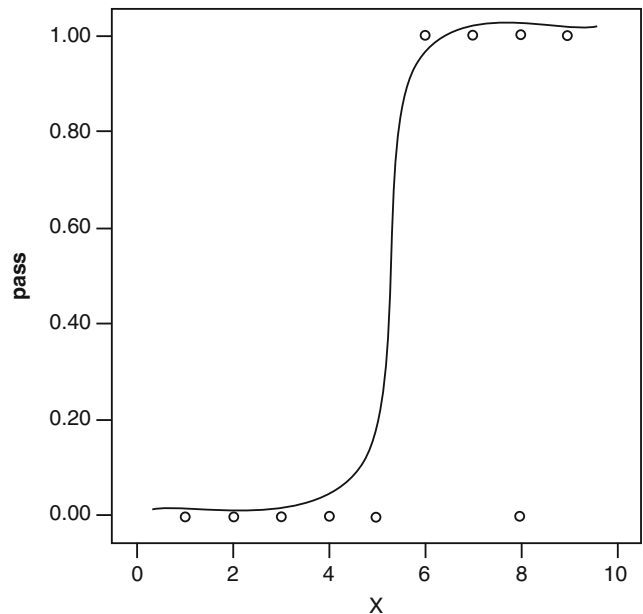
regression, we predict the *odds* or *log-odds* (logit) that a certain condition will or will not happen. **Odds** range from 0 to infinity and are a ratio of the chance of an event (p) divided by the chance of the event not happening, that is, $p/(1-p)$. **Log-odds** (logits) are transformed odds, $\ln[p/(1-p)]$, and range from negative to positive infinity. The relationship predicting probability using x follows an S-shaped curve as shown in Figure 3. The shape of curve above is called a “logistic curve.” This is defined as $p(y_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \epsilon_i)}{1 + \exp(\beta_0 + \beta_1 x_i + \epsilon_i)}$. In this logistic regression, the value predicted by the equation is a log-odds or logit. This means when we run logistic regression and get coefficients, the values the equation produces are logits. Odds is computed as $\exp(\text{logit})$, and probability is computed as $\frac{\exp(\text{logit})}{1 + \exp(\text{logit})}$. Another model used to predict binary outcome is the **probit model**, with the difference between logistic and probit models lying in the assumption about the distribution of errors: while the logit model assumes standard logistic distribution of errors, probit model assumes normal distribution of errors (Chumney & Simpson 2006). Despite the difference in assumption, the predictive results using these two models are very similar. When the outcome variable has multiple

categories, **multinomial logistic regression** or **ordered logistic regression** should be implemented depending on whether the dependent variable is nominal or ordinal.

Regression in Big Data

Due to the advanced technologies that have been increasingly used in data collection and the vast amount of user-generated data, the amount of data will continue to increase at a rapid pace, along with a growing accumulation of scholarly works. The explosion of knowledge makes big data one of new research frontiers with an extensive number of application areas affected by big data, such as public health, social science, finance, geography, and so on. The high volume and complex structure of big data bring statisticians both opportunities and challenges. Generally speaking, big data is a collection of large-scale and complex data sets that are difficult to be processed and analyzed using traditional data analytic tools. Inspired by the advent of machine learning and other disciplines, **statistical learning** has emerged as a new subfield in statistics, including supervised and unsupervised statistical

Regression,
Fig. 3 Logistic regression
models



learning (James, Witten, Hastie, & Tibshirani, 2013). **Supervised statistical learning** refers to a set of approaches for estimating the function f based on the observed data points, to understand the relationship between Y and $X = (X_1, X_2, \dots, X_p)$, which can be represented as $Y = f(X) + \varepsilon$. Since the two main purposes for the estimation are to make **prediction** and **inference**, which regression modeling is widely used for, many classical statistical learning methods use regression models, such as linear, nonlinear, and logistic regression, with the selection of specific regression model based on research question and data structure. In contrast, for **unsupervised statistical learning**, there is no response variable to predict for every observation that can supervise our analysis (James et al. 2013). Additionally, more methods have been recently developed, such as **Bayesian** and **Markov chain Monte Carlo (MCMC)**. Bayesian approach, distinct from the frequentist approach, treats model parameters as random and models them via distributions. MCMC is statistical sampling investigations that involve sample data generation to obtain empirical sampling distributions based on constructing a Markov chain that has the desired distribution (Bandalos & Leite 2013).

Cross-References

- [Data Mining](#)
- [Data Mining Algorithms](#)
- [Machine Learning](#)
- [Statistics](#)

Further Reading

- Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 625-666). Charlotte, NC: Information Age Publishing.
- Chumney, E. C., & Simpson, K. N. (2006). *Methods and designs for outcomes research*. Bethesda, MD: ASHP.
- Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York, NY: Springer.
- Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. New York, NY: Routledge.

Regulation

Christopher Round

George Mason University, Fairfax, VA, USA

Booz Allen Hamilton, Inc., McLean, VA, USA

Synonyms

[Governance instrument](#); [Policy](#); [Rule](#)

Regulations may be issued for different reasons. Regulations may be issued to address collective desires, diversify or limit social experiences, perform interest group transfers, or address market failures. Government regulations can be used to address market failures such as negative externalities. This allows for the validation of the base assumptions economists believe are necessary to establish for a free market to operate. Regulations may also be used to codify behavior or norms that are deemed by the organization issuing the regulation as beneficial.

Regulations are designed by their issuing body to address a target issue. There is a wide variety in their potential design. Regulations can be direct or indirect, downstream or upstream, and can be influenced by outside concerns such as who will be indirectly impacted. Direct regulations are aimed to address the issue at hand by fitting the regulation as closely as possible to the issue. For example, a direct regulation on pollution emissions would issue some form of limit on the pollution release (e.g., limiting greenhouse gas emissions to level X from a power plant). An indirect regulation seeks to address an issue by impacting a related issue. For example, a regulation improving gas mileage for vehicles would indirectly reduce greenhouse gas emissions. Regulations can be

downstream or upstream (Kolstad 2010). An upstream regulation seeks to influence decision-making in relation to an issue by affecting the source of the issue (typically the producer). A downstream regulation seeks to influence an issue by changing the behavior of individuals or organizations who have influence on the originator of the issue (typically the consumers). For example, an upstream regulation on greenhouse gas emissions may impact fossil fuel production. A downstream regulation could be a limitation on fossil fuel purchases by consumers. Indirect factors such as the burden of cost of the regulation and who bears it will influence questions of regulation design (Kolstad 2010).

Regulations can take different forms based on the philosophical approach and direct and indirect considerations of decision-makers (Cole and Grossman 1999; Kolstad 2010). Command and control regulations provide a prescription for choices by the regulated community, such as a limit on the number of taxi medallions or a nightly curfew (Cole and Grossman 1999; Kolstad 2010). Technical specification regulations are a form of command and control regulation dictating what technology may be used for a product (Cole and Grossman 1999; Kolstad 2010). Regulations may take the form of market mechanisms, such as a penalty or subsidy to influence the behavior of actors in a market (Cole and Grossman 1999; Kolstad 2010).

Regulations may be issued with an ulterior agenda than to serve the general population represented by a governing body. Regulatory capture is a diagnosis of a regulating body in which the regulating body is serving a special interest over the interests of the wider population it impacts and is a form of corruption (Carpenter and Moss 2014a; Levine and Forrence 1990). This can be done to entrench the power or economic interests of a specific group, to manipulate markets, or to weaken or strengthen regulations to benefit a specific interest.

Regulatory capture can take two forms: cultural and material capture. Cultural capture occurs when the norms and preferences of the regulated community over time permeate into the regulated body and influence it to make decisions

considered friendly by special interests within the regulated community (Carpenter & Moss, 2014a). Material capture is a form of principle-agent interaction where a special interest material regulatory capture can only be diagnosed if there is demonstrable proof that a regulation issued originated from a third party (Carpenter and Moss 2014a, b; Levine and Forrence 1990; Susan Webb Yackee 2014).

Big data itself is subject to multiple regulations depending on the information it contains and the location of the entity responsible for it. Data containing personally identifiable information (PII) is of particular concern, especially if it contains information that individuals may wish to keep private such as their medical history. In Europe, big data is regulated under the General Data Protection Regulation (GDPR) (European Parliament and Council 2018). Within the USA, data is regulated by entities at different levels of governance with no single overarching legal overview (Chabinsky and Pittman 2019). Thus, individuals and organizations utilizing big data in the USA will need to consult with local rules and subject matter-based regulations in order to ensure compliance. At the federal level, the US Federal Trade Commission is tasked with enforcing federal privacy and data protection regulations. Specific types of data are regulated under different legal authorities such as medical data which is regulated under the Health Insurance Portability and Accountability Act. Major state-level laws include the California Consumer Privacy Act.

Further Reading

- Carpenter, D. P., & Moss, D. A. (2014a). Introduction. In D. P. Carpenter & D. A. Moss (Eds.), *Preventing regulatory capture: Special interest influence and how to limit it* (pp. 1–22). Cambridge: Cambridge University Press.
- Carpenter, D. P., & Moss, D. A. (Eds.). (2014b). *Preventing regulatory capture: Special interest influence and how to limit it*. Cambridge: Cambridge University Press.
- Chabinsky, S., & Pittman, F. P. (2019, March 7). *USA Data Protection 2019* (United Kingdom) [Text]. International Comparative Legal Guides International

- Business Reports; Global Legal Group. <https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa>.
- Cole, D. H., & Grossman, P. Z. (1999). When is command-and-control efficient? Institutions, Technology, and the Comparative Efficiency of Alternative Regulatory Regimes for Environmental Protection. *Articles by Maurer Faculty*, Paper 590.
- European Parliament and Council. (2018). REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL – of 27 April 2016 – On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union Law*, 119, 1–80.
- Kolstad, C. (2010). *Environmental Economics* (2nd ed.). New York: Oxford University Press.
- Levine, M. E., & Forrence, J. L. (1990). Regulatory capture, public interest, and the public agenda: Toward a synthesis. *Journal of Law, Economics, and Organization*, 6, 167–198.
- Merriam-Webster. (2018). Definition of REGULATION. Merriam-Webster.Com. Retrieved July 31, 2018, from <https://www.merriam-webster.com/dictionary/regulation>.
- Susan Webb Yackee. (2014). Reconsidering Agency Capture During Regulatory Policymaking. In D. P. Carpenter & D. A. Moss (Eds.), *Preventing regulatory capture: Special interest influence and how to limit it* (pp. 292–325). Cambridge University Press. <https://www.tobinproject.org/sites/tobinproject.org/files/assets/Kwak%20-%20Cultural%20Capture%20and%20the%20Financial%20Crisis.pdf>.
- Visseren-Hamakers, I. J. (2015). Integrative environmental governance: Enhancing governance in the era of synergies. *Current Opinion in Environmental Sustainability*, 14, 136–143. <https://doi.org/10.1016/j.cosust.2015.05.008>.

Relational Data Analytics

► Link/Graph Mining

Religion

Matthew Pittman and Kim Sheehan
School of Journalism & Communication,
University of Oregon, Eugene, OR, USA

In his work on the changing nature of religion in our modern mediated age, Stewart Hoover notes

that religion today is much more commodified, therapeutic, public, and personalized than it has been for most of history. He also notes that, because media are coming together to create an environment in which our personal projects of identity, meaning, and self are worked out, religion and media are actually converging. As more people around the globe obtain devices capable of accessing the Internet, their everyday religious practices are leaving digital traces for interested companies and institutions to pick up on. The age of big data is usually thought to affect institutions like education, mass media, or law, but religion is undergoing dynamic shifts as well.

Though religious practice was thought to be in decline through the end of the twentieth century, there has been a resurgence of interest through the beginning of the 21st. A Google NGram viewer (which keeps track of a word's frequency in published books and general literature over time) shows that “data” surpassed “God” for the first time in 1973. Yet, by about 2004, God once again overtook data (and its synonym “information”), indicating that despite incredible scientific and technological advances, people still wrestle with spiritual or existential matters.

While the term “big data” seems commonplace now, it is a fairly recent development. Several researchers and authors claim to have coined the term, but its modern usage took off in the mid-1990s and only really became mainstream in 2012 when the White House and the Davos World Economic Forum identified it as a serious issue worth tackling. Big data is a broad term, but generally has two main precepts: humans are now producing information at an unprecedented rate, and new methods of analysis are needed to make sense of that information. Religious practices are changing in both of these areas. Faith-based activity is creating new data streams even as churches, temples, and mosques are figuring out what to do with all that data. On an institutional level, the age of big data is giving religious groups new ways to learn about the individuals who adhere to their teachings. On an individual level, technology is changing how people across the globe learn about, discuss, and practice their faiths.

Institutional Religion

It is now common for religious institutions to use digital technology to reach their believers. Like any other business or group that needs members to survive, most seek to utilize or leverage new devices and trends into opportunities to strengthen existing members or recruit potential new ones. Of course, depending on a religion's stance toward culture, they may (like the Amish) eschew some technology. However, for most mosques, churches, and synagogues, it has become standard for each to have its own website or Facebook page. Email newsletters and Twitter accounts feeds have replaced traditional newsletters and event reminders.

New opportunities are constantly emerging that create novel space for leaders to engage practitioners. Religious leaders can communicate directly with followers through social media, adding a personal touch to digital messages, which can sometimes feel distant or cold. Rabbi Schmuley Boteach, "America's Rabbi," has 29 best-selling books but often communicates daily through his Twitter account, which has over a hundred thousand followers. On the flip side, people can thoroughly vet potential religious leaders or organizations before committing to them. If concerned that a particular group's ideology might not align with one's own, a quick Internet search or trip to the group's website should identify any potential conflicts. In this way, providing data about their identity and beliefs helps religious groups differentiate themselves.

In a sense, big data makes it possible for religious institutions to function more like – and take their cues from – commercial enterprises. Tracking streams of information about its followers can help religious groups be more in tune with the wants and needs of these "customers." Some religious organizations implement the retail practice of "tweets and seats": by ensuring that members always have available places to sit, rest, or hang out, and that wifi (wireless Internet connectivity) is always accessible, they hope to keep people present and engaged. Not all congregations embrace this

change, but the clear cultural trend is toward ubiquitous smart phone connectivity. Religious groups that take advantage of this may provide several benefits to their followers: members could immediately identify and download any worship music being played; interested members could look up information about a local religious leader; members could sign up for events and groups as they are announced in the service; or those using online scripture software can access texts and take notes. There are just a few possibilities.

There are other ways religious groups can harness big data. Some churches have begun analyzing liturgies to assess and track length and content over time. For example, a dip in attendance during a given month might be linked to the sermons being 40% longer in that same time frame. Many churches make their budgets available to members for the sake of transparency, and in a digital age it is not difficult to create financial records that are clear and accessible to laypeople. Finally, learning from a congregant's social media profiles and personal information, a church might remind a parishioner of her daughter's upcoming birthday, the approaching deadline for an application to a family retreat, or when other congregants are attending a sporting event of which she is a fan. The risk of overstepping boundaries is real and, just like with Facebook or similar entities, privacy settings should be negotiated beforehand.

As with other commercial entities, religious institutions utilizing big data must learn to differentiate information they need from information they don't. The sheer volume of available data makes distinguishing desired signal from irrelevant noise an increasingly important task. Random correlations may lead to false positive causation. A mosque may benefit from learning that members with the highest income are not actually its biggest givers, or testing for a relationship between how far away its members live and how often they attend. Each religious group must determine how big data may or may not benefit its operation in any given endeavor, and the opportunities are growing.

Individual Religion

The everyday practice of religion is becoming easier to track as it increasingly utilizes digital technology. A religious individual's personal blog, Twitter feed, Facebook profile keep a record of his or her activity or beliefs, making it relatively easy for any interested entity to track online behavior over time. Producers and advertisers use this data to promote products, events, or websites to people who might be interested. Currently companies like Amazon have more incentive than, say, a local synagogue in keeping tabs on what websites one visits, but the potential exists for religious groups to access the same data that Facebook, Amazon, Google, etc. already possess.

Culturally progressive religious groups anticipate mutually beneficial scenarios: they provide a data service that benefits personal spiritual growth, and in turn the members generate fields of data that are of great value to the group. A Sikh coalition created the FlyRights app in 2012 to help with quick reporting of discriminatory TSA profiling while travelling. The Muslim's Prayer Times app provides a compass, calendar (with moon phases), and reminders for Muslims about when and in what direction to pray. Apple's app store has also had to ban other apps from fringe religious groups or individuals for being too irreverent or offensive.

The most popular religious app to date simply provides access to scripture. In 2008 LifeChurch.tv launched "the Bible app," also called YouVersion, and it currently has over 151 million installations worldwide on smartphones and tablets. Users can access scripture (in over 90 different translations) while online or download it for access offline. An audio recording of each chapter being read aloud can also be downloaded for some of the translations. A user can search through scripture by keyword, phrase, or book of the Bible, or there are reading plans of varying levels of intensity and access to related videos or movies. A "live" option lets users search out churches and events in surrounding geographic areas, and a sharing option lets users promote the app, post to social media what they have read, or

share personal notes directly to friends. The digital highlights or notes made, even when using the app offline, will later upload to one's account and remain in one's digital "bible" permanently.

All this activity has generated copious amounts of data for YouVersion's producers. In addition to using the data to improve their product they also released it to the public. This kind of insight into the personal religious behavior of so many individuals is unprecedented. With over a billion opens and/or uses, YouVersion statistically proved several phenomena. The data demonstrated the most frequent activity for users is looking up a favorite verse for encouragement. Despite the stereotype of shirtless men at football games, the most popular verse was not John 3:16, but Philipians 4:13: "I can do all things through him who gives me strength." Religious adherents have always claimed that their faith gives them strength and hope, but big data has now provided a brief insight into one concrete way this actually happens.

The YouVersion data also reveal that people used the bible to make a point in social media. Verses were sought out and shared in an attempt to support views on marriage equality, gender roles, or other divisive topics. Tracking how individuals claim to have their beliefs supported by scripture may help religious leaders learn more about how these beliefs are formed, how they change over time, and which interpretations of scripture are most influential. Finally, YouVersion data reveal that Christian users like verses with simple messages, but chapters with profound ideas. Verses are easier to memorize when they are short and unique, but when engaging in sustained reading, believers prefer chapters with more depth. Whether large data sets confirm suspicions or shatter expectations, they continue to change the way religion is practiced and understood.

Numerous or Numinous

In the past, spiritual individuals had a few religions to choose from, but the globalizing force of technology has dramatically increased the

available options. While the three big monotheisms (Christianity, Judaism, and Islam) and pan/polytheisms (Hinduism and Buddhism) are still the most popular, the Internet has made it possible for people of any faith, sect, or belief to find each other and validate their practice. Though pluralism is not embraced in every culture, there is at least increasing awareness of the many ways religion is practiced across the globe.

Additionally, more and more people are identifying themselves as “spiritual but not religious,” indicating a desire to seek out spiritual experiences and questions outside the confines of a traditional religion. Thus for discursive activities centered on religion, Daniel Stout advocates the use of another term in addition to “religion”: numinous. Because “religious” can have negative or limiting connotations, looking for the “numinous” in cultural texts or trends can broaden the search for and dialogue about a given topic. To be numinous, something must meet several criteria: stir deep feeling (affect), spark belief (cognition), include ritual (behavior), and be done with fellow believers (community). This four-part framework is a helpful tool for identification of numinous activity in a society where it once might have been labeled “religious.”

By this definition, the Internet (in general) and entertainment media (in particular) all contain numinous potential. The flexibility of the Internet makes it relevant to the needs of most; while authority of some of its sources can be dubious, the ease of social networking and multi-mediated experiences provides all the elements of traditional religion (community, ritual, belief, feeling). Entertainment media, which produce at least as much data as – and may be indistinguishable from – religious media, emphasize universal truths through storytelling. The growing opportunities of big data (and its practical analysis) will continue to offer for those who engage in numinous and religious behavior.

Cross-References

- [Data Monetization](#)
- [Entertainment](#)

Further Reading

- Campbell, H. A. (Ed.). (2012). *Digital religion: Understanding religious practice in new media worlds*. Abingdon: Routledge.
- Hjarvard, S. (2008). The mediatization of religion: A theory of the media as agents of religious change. *Northern Lights: Film & Media Studies Yearbook*, 6(1), 9–26.
- Hoover, S. M., & Lundby, K. (Eds.). (1997). *Rethinking media, religion, and culture* (Vol. 23). Thousand Oaks: Sage.
- Kuruvilla, C. Religious mobile apps changing the faith-based landscape in America. Retrieved from <http://www.nydailynews.com/news/national/gutenberg-moment-mobile-apps-changing-america-religious-landscape-article-1.1527004>. Accessed Sep 2014.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- Taylor, B. (2008). *Entertainment theology (cultural exegesis): New-edge spirituality in a digital democracy*. Baker Books.

Risk Analysis

Jonathan Z. Bakdash
Human Research and Engineering Directorate,
U.S. Army Research Laboratory, Aberdeen
Proving Ground, MD, USA

Definition and Introduction

Society is becoming increasingly interconnected with networks linking people, the environment, information, and technology. This rising complexity is a challenge for risk analysis. Risk analysis is the identification and evaluation of the probability of an adverse outcome, its associated risk factors, and the potential impact if that outcome occurs. Successfully modeling risk within interdependent and complex systems requires access to considerably more data than traditional, simple risk models. The increasing availability of big data offers enormous promise for improving risk analysis through more detailed, comprehensive, faster, and accurate predictions of risks and their impacts than small data alone.

However, risk analysis is not purely a computational challenge that can be solved by more data. Big data does not eliminate the importance of data quality and modeling assumptions; it is not necessarily a replacement for small data. Furthermore, traditional risk analysis methods typically underestimate the probability and impact of risks (e.g., terrorist attacks, power failures, and natural disasters such as hurricanes) because normal data and independent observations are assumed. Traditional methods also typically do not account for cascading failures, which are not uncommon in complex systems. For example, a hurricane may cause a power failure, which in turn results in flooding.

The blessing and curse of risk analysis with big data are illustrated by the example of Google Flu Trends (GFT). Initially, it was highly successful in estimating flu rates in real time, but over time it became inaccurate due to external factors, lack of continued validation, and incorrect modeling assumptions.

Interdependencies

Globalization and advances in technology have led to highly networked and interdependent social, economic, political, natural, and technological systems (Helbing 2013). Strong interdependencies are potentially dangerous because small or gradual changes in a single system can cause cascading failures throughout multiple systems. For example, climate change is associated with food availability, food availability with economic disparity, and economic disparity with war. In interconnected systems, risks often spread quickly in a cascading process, so early detection and mitigation of risks is critical to stopping failures before they become uncontrollable. Helbing (2013) contends that big data is necessary to model risks in interconnected and complex systems: Capturing interdependent dynamics and other properties of systems requires vast amounts of heterogeneous data over space and time.

Interdependencies are also critical to risk analysis because even when risks are mitigated, they

may still cause amplifying negative effects because of human risk perception. Perceived risk is the public social, political, and economic impacts of unrealized (and realized) risks. An example of the impact of a perceived risk is the nuclear power accident at Three-Mile Island. In this accident, minimal radiation was released so the real risk was mitigated. Nevertheless, the near miss of a nuclear meltdown had immense social and political consequences that continue to negatively impact the nuclear power industry in the United States. The realized consequences of perceived risk mean that “real” risk should not necessarily be separated from “perceived” risk.

Data: Quality and Sources

Many of the analysis challenges for big data are not unique but are pertinent to analysis of all data (Lazer et al. 2014). Regardless of the size of the dataset, it is important for analysts and policymakers to understand how, why, when, and where the data were collected and what the data contain and do not contain. Big data may be “poor data” because rules, causality, and outcomes are far less clear compared to small data.

More specifically, Vose (2008) describes the quality of data characteristics for risk analysis. The highest quality data are obtained using a large sample of direct and independent measurements collected and analyzed using established best practices over a long period of time and continually validated to correct data for errors. The second highest quality data use proxy measures, a widely used method for collection, analysis, and some validation. Other characteristics of decreasing data quality are: A smaller sample of objective data, agreement among multiple experts, a single expert opinion, and is weakest with speculation. While there may be some situations in which expert opinions are the only data source, general findings indicate this type of data has poor predictive accuracy. Additional reasons to question experts are situations or systems with a large number of unknown factors and potentially catastrophic impacts for erroneous estimations. Big data can be an improvement over small data

and one or several expert opinions. However, volume is not necessarily the same as quality. Multidimensional aspects of data quality, whether the data are big or small, should always be considered.

Risk Analysis Methods

Vose (2008) explains the general techniques for conducting risk analysis. A common, descriptive method for risk analysis is Probability-Impact (P-I). P-I is the probability of a risk occurring multiplied by the impact of the risk if it materializes: $Probability \times Impact = Weighted Risk$. All values may be either qualitative (e.g., low, medium, and high likelihood or severity) or quantitative (e.g., 10% or one million dollars). The *Probability* may be a single value or multiple values such as a distribution of probabilities. The *Impact* may also be a single value or multiple values and is usually expressed as money. A similar weighted model to P-I, $Threat \times Vulnerability \times Consequence = Risk$, is frequently used in risk analysis. However, a significant weakness with P-I and related models with fixed values is that they tend to systematically underestimate the probability and impact of rare events that are interconnected, such as natural hazards (e.g., floods), protection of infrastructure (e.g., power grid), and terrorist attacks. Nevertheless, the P-I method can be effective for quick risk assessments.

Probabilistic Risk Assessment

P-I is a foundation for Probabilistic Risk Assessment (PRA), an evaluation of the probabilities for multiple potential risks and their respective impacts. The US Army's standardized risk matrix is an example of qualitative PRA, see Fig. 1 (also see Level 5 of risk analysis below).

The risk matrix is constructed by:

- Step 1:* Identifying possible hazards (i.e., potential risks)
- Step 2:* Estimating the probabilities and impacts of each risk and using the P-Is to categorize weighted risk

Risk analysis informs risk reduction, but they are not one and the same. After the risk matrix is constructed, appropriate risk tolerance and mitigation strategies are considered. The last step is ongoing supervision and evaluation of risk as conditions and information change, updating the risk matrix as needed, and provided feedback to improve the accuracy of future risk matrix.

Other widely used techniques include inferential statistical tests (e.g., regression) and the more comprehensive approach of what-if data simulations, which are also used in catastrophe modeling. Big data may improve the accuracy of probability and impact estimates, particularly the upper bounds in catastrophe modeling, leading to more accurate risk analysis.

From a statistical perspective, uncertainty and variability tend to be interchangeable. If uncertainty can be attributed to random variability, there is no distinction. However, in risk analysis, uncertainty can arise from incomplete knowledge (Paté-Cornell 1996). Uncertainty in risk may be due to a lack of data (particularly for rare events), not knowing relevant risks and/or impacts and unknown interdependencies among risks and/or impacts.

Levels of Risk Analysis

There are six levels for understanding uncertainty, ranging from qualitative identification of risk factors (Level 0) to multiple risk curves constructed using different PRAs (Level 5) (Paté-Cornell 1996). Big data are relevant to Level 2 and beyond. The specific levels are as follows (adapted Paté-Cornell 1996):

- Level 0: Identification of a hazard or failure modes.* Level 0 is primarily qualitative. For example, does exposure to a chemical increase the risk of cancer?
- Level 1: Worst case.* Level 1 is also qualitative, with no explicit probability. For example, if individuals are exposed to a cancer-causing chemical, what is the highest number that could develop cancer?
- Level 2: Quasi-worst case (probabilistic upper-bound).* Level 2 introduces subjective estimation of probability based on reasonable expectation(s). Using the example from Level 1, this

Table 3-3.
Standardized Army risk matrix

		Probability (expected frequency)				
		Frequent: Continuous, regular, or inevitable occurrences	Likely: Several or numerous occurrences	Occasional: Sporadic or intermittent occurrences	Seldom: infrequent occurrences	Unlikely: Possible occurrences but improbable
Severity (expected consequence)		A	B	C	D	E
Catastrophic: Death, unacceptable loss or damage, mission failure, or unit readiness eliminated	I	EH	EH	H	H	M
Critical: Severe injury, illness, loss, or damage; significantly degraded unit readiness or mission capability	II	EH	H	H	M	L
Moderate: Minor injury, illness, loss, or damage; degraded unit readiness or mission capability	III	H	M	M	L	L
Negligible: Minimal injury, loss, or damage; little or no impact to unit readiness or mission capability	IV	M	L	L	L	L

Legend for Table 3-3:
EH – extremely high risk
H – high risk
L – low risk
M – medium risk

Table 3-4.
Risk matrix codes and descriptions

Symbol	Risk Assessment Code (RAC)	Description
EH	1	Extremely High
H	2	High
M	3	Medium
L	4	Low

Risk Analysis, Fig. 1 Risk analysis (Source: Safety Risk Management, Pamphlet 385-30 (Headquarters, Department of the Army, 2014, p. 8): www.apd.army.mil/pdffiles/p385_30.pdf)

could be the 95th percentile for the number of individuals developing cancer.

Level 3: Best and central estimates. Rather than a worst case, Level 3 aims to model the most likely impact using central values (e.g., mean or median).

Level 4: Single-curve PRA. Previous levels were point estimates of risk; Level 4 is a type of PRA. For example, what is the number of individuals that will develop cancer across a probability distribution?

Level 5: Multiple-curve PRA. Level 5 has more than one probabilistic risk curve. Using the cancer risk example, different probabilities from distinct data can be represented using multiple curves, which are then combined using the average or another measure. A generic example of Level 5, for qualitative values, was illustrated with the above risk matrix. When implemented quantitatively, Level 5 is similar to what-if simulations in catastrophe modeling.

Catastrophe Modeling

Big data may improve risk analysis at Level 2 and above but may be particularly informative for modeling multiple risks at Level 5. Using catastrophe modeling, big data can allow for a more comprehensive analysis of the combinations of P-Is while taking into account interdependences among systems. Catastrophe modeling involves running a large number of simulations to construct a landscape of risk probabilities and their impacts for events such as terrorist attacks, natural disasters, and economic failures. Insurance, finance, other industries, and governments are increasingly relying on big data to identify and mitigate interconnected risks using catastrophe modeling.

Beiser (2008) describes the high level of data detail in catastrophe modeling. For risk analysis of a terrorist attack in a particular location, interconnected variables taken into account may include the proximity to high-profile targets (e.g., government buildings, airports, and landmarks), the city, and details of the surrounding buildings (e.g., construction materials), as well as the potential size and impact of an attack. Simulations are run under different assumptions, including the likelihood of acquiring materials to carry out a particular type of attack (e.g., a conventional bomb versus a biological weapon) and the probability of detecting the acquisition of such materials. Big data is informative for the wide range of possible outcomes and their impacts in terms of projected loss of life and property damage. However, risk analysis methods are only as good as their assumptions, regardless of the amount of data.

Assumptions: Cascading Failures

Even with big data, risk analysis can be flawed due to inappropriate model assumptions. In the case of Hurricane Katrina, the model assumptions for a Category 3 hurricane did specify a large, slow-moving storm system with heavy rainfall nor did they account for the interdependencies in infrastructure systems. This storm caused early loss of electrical power, so many of the pumping stations for levees could not operate. Consequently, water overflowed, causing breaches, resulting in widespread flooding. Because of

cascading effects in interconnected systems, risk probabilities and impacts are generally far greater than in independent systems and therefore will be substantially underestimated when incorrectly treated as independent.

Right Then Wrong: Google Flu Trends

GFT is an example of both success and failure for risk analysis using big data. The information provided by an effective disease surveillance tool can help mitigate disease spread by reducing illnesses and fatalities. Initially, GFT was a successful real-time predictor of flu prevalence, but over time, it becomes inaccurate. This is because the model assumptions did not hold over time, validation with small data was not on-going, and it lacked transparency. GFT used a data-mining approach to estimate real-time flu rates: Hundreds of millions of possible models were tested to determine the best fit of millions of Google searches to traditional weekly surveillance data. The traditional weekly surveillance data consisted of the proportion of reported doctor visits for flu-like systems. At first, GFT was a timely and accurate predictor of flu prevalence, but it began to produce systematic overestimates, sometimes by a factor of two or greater compared with the gold-standard of traditional surveillance data. The erroneous estimates from GFT resulted from a lack of continued validation, thus assuming relevant search terms only changed as a result of flu symptoms and transparency in the data and algorithms used.

Lazer et al. (2014) called the inaccuracy of GFT a parable for big data, highlighting several key points. First, a key cause for the misestimates was that the algorithm assumed that influences on search patterns were the same over time and primarily driven by the onset of flu symptoms. In reality, searches were likely influenced by external events such as media reporting of a possible flu pandemic, seasonal increases in searches for cold symptoms that were similar to flu symptoms, and the introduction of suggestions in Google Search. Therefore, GFT wrongly assumed the data were stationary (i.e., no trends or changes in the mean and variance of data over time). Second,

Google did not provide sufficient information for understanding the analysis, such as all selected search terms and access to the raw data and algorithms. Third, big data is not necessarily a replacement for small data. Critically, the increased volume of data does not necessarily make it the highest quality source. Despite these issues, GFT was at the second highest level of data quality using criteria from Vose (2008) because GFT initially used:

1. Proxy measures: search terms originally correlated with local flu reports over a finite period of time
2. A common method: search terms used for Internet advertising, disease surveillance was novel (with limited validation)

In the case of GFT, the combination of big and small data, by continuously recalibrating the algorithms for the big data using the small (surveillance) data, would have been much more accurate than either alone. Moreover, big data can make powerful predictions that are impossible with small data alone. For example, GFT could provide estimates of flu prevalence in local geographic areas using detailed spatial and temporal information from searches; this would be impossible with only the aggregated traditional surveillance data.

Conclusions

Similar to GFT, many popular techniques for analyzing big data use data mining to automatically uncover hidden structures. Data mining techniques are valuable for identifying patterns in big data but should be interpreted with caution. The dimensions of big data do not obviate considerations of data quality, the need for continuous validation, and the importance of modeling assumptions (e.g., non-normality, non-stationarity, and non-independence). While big data has enormous potential to improve the accuracy and insights of risk analysis, particularly for interdependent systems, it is not necessarily a replacement for small data.

Cross-References

- [Complex Networks](#)
- [Financial Data and Trend Prediction](#)
- [Google Flu](#)
- [“Small” Data](#)

References

- Beiser, V. (2008). Pricing terrorism: Insurers gauge risks, costs, *Wired*. Permanent link: http://web.archive.org/save/_embed/http://www.wired.com/2008/06/pb-terror-ism/.
- Helbing, D. (2013). Globally networked risks and how to respond. *Nature*, 497(7447), 51–59. doi:10.1038/nature12047.
- Lazer, D. M., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, 343(6176), 1203–1206. doi:10.1126/science.1248506.
- Paté-Cornell, M. E. (1996). Uncertainties in risk analysis: Six levels of treatment. *Reliability Engineering & System Safety*, 54(2), 95–111. doi:10.1016/S0951-8320(96)00067-1.
- Vose, D. (2008). *Risk analysis: A quantitative guide* (3rd ed.). West Sussex: Wiley.

R-Programming

Anamaria Berea

Department of Computational and Data Sciences,
George Mason University, Fairfax, VA, USA
Center for Complexity in Business, University of
Maryland, College Park, MD, USA

R is an open-source software programming language and software environment for statistical computing and graphics that is based on object-oriented programming (R Core Team 2016). Originally, R was an implementation of the S-programming language and it has been extended with various packages, functions, and extensions. There is a large R-community of users and developers who are continuously contributing to the development of R (Muenchen 2012). R is available under the GNU General Public License. One of the most used online forums of the R-community is

Stack Overflow, and one of the most used online blogs is r-bloggers (<http://www.r-bloggers.com>).

As of May 2017, there were included more than 10,500 additional packages and 120,000 functions with the installation of R. These are available at the Comprehensive R Archive Network (CRAN).

Arguably, R-language has become one of the most important tools for computational statistics, visualization, and data science. Worldwide, millions of statisticians and data scientists use R to solve their most challenging problems in fields ranging from computational biology to quantitative marketing (Matloff 2011).

This software is easily accessible, and anyone can use it as it is open source (there is no purchasing fee). R can be used with textual code scripts as well as inside an environment (RStudio). R code and scripts can be written to analyze the data or to fully implement simulations. In other words, R can handle computational jobs from the simplest data analyses, such as showing ranges and simple statistics of the data (minimum and maximum values), to complex models, such as ARIMA, Bayesian networks, Monte Carlo simulations, and agent-based simulations.

Once the program is created and used with data, various graphic displays can be created quite easily. Once you become familiar with this programming language, R is an easy tool to use, but if not, it takes a short time to learn how to use, as currently there are also many online tutorials. Additionally, there are many books that exist that can explain how to use this software.

R and Big Data

R is a very powerful tool for analyzing large datasets. In one code run of R, datasets as large as tens of millions of data points can be analyzed and crunched within a reasonable time on a personal computer. For truly Big Data, that requires parallel or distributed computing, R can be used with a series of packages called pdbR (Raim 2013). In this case, data is analyzed in batches.

For streaming data, which requires different data architectures, cleaning, and collection

processes than batch data, R can be used for data analytics and visualizations as R-server. The R-server can be connected to other databases and run the analytics and visualizations either through direct ODBC connections or through reading APIs. The R-server can be set up either on AWS (Amazon web Services) or can be bought as an enterprise solution from Microsoft.

Comparison with Other Statistical Software

SPSS

SPSS is a well-known statistical software that has been used as a business solution for companies. SPSS user interface looks quite similar to Microsoft Excel, which is widely known by most professionals. They can therefore easily apply their knowledge to this new program. Additionally, the graphs and visualizations can be easily customized and are more visually appealing. The trade-off is that SPSS cannot do complex analyses and there is a limitation to the size of the data that can be analyzed in one batch. This is a commercial solution, not open source.

SAS

For advanced analytics, SAS program has been one of the most widely used. It is quite similar to R, yet it is not open source and not open to the public. SAS is more difficult to learn than both SPSS and Stata, but can run more complicated analyses than both of them. On another hand, it is easier to use than R, but just like SPSS, it is not suitable for Big Data or complex, noisy data. SAS is also hard to implement in data streaming environments.

Stata

This is a command-based software in which the user writes code to produce analytical results similar to R. It is widely used by researchers and professionals, as it creates impressive-looking output tables. Different versions of Stata can be purchased for different needs and budgets. It is easier to learn than R and SAS, but much more complicated than SPSS. There is a journal called

The Stata Journal which releases information about work that has been done with Stata and how to use the program more efficiently. Additionally, Stata holds an annual conference in which developers meet and present. On another hand, Stata is not suitable for large and noisy datasets either, as the cleaning of the data is much more difficult to do using Stata than using R.

Python

Python is considered the closest competitor to R regarding the analysis and visualization of Big Data and of complex datasets. Python was developed for programmers, while R was developed with the statisticians and statistics in mind. While Python is a general purpose language and has an easier syntax than R, R is more often praised for its features on data visualization and complex statistical analyses. Python is more focused on code readability and transferability, while R is specific for graphical models and data analysis. Both languages can be used to perform more complex analyses, such as natural language programming or geospatial analyses, but Python scales up better than R for large, complex data architectures. R is being used more by statisticians and researchers, while Python is being used more by engineers and computer programmers. Due to their different syntax styles, R is more difficult to learn in the beginning than Python, but after the learning curve is crossed, R can be easier to use than Python.

R Syntax and Use

R uses command-line scripting and one of the “trademarks” of the R syntax is the use of the inverse arrow $\leftarrow$ for defining objects inside the code – the assignment operator (*example*: $x \leftarrow \text{sample}(1:10, 1)$). For programmers from other languages, the syntax may look peculiar at first, but it is an easy-to-learn syntax, with plenty of tutorials and support online (Cook 2017). Another peculiarity is the way R uses the “\$” operator to call variables inside a data set, similar to the way other languages use “.” (the dot). On

another hand, functions are defined in a similar way to JavaScript (Cook 2017).

More than Statistics Programming

R programming is not only a statistical or Big Data type of programming language. Due to the development of many packages and the versatility given by functional programming in general, R can be successfully used for text mining, geospatial visualizations, artistic visualizations, and even agent-based modeling. For example, R has a package {tm} that can be used for text mining and the analysis of literary corpuses. Another package, {topicmodeling}, can be used as a natural language processing technique to discover topics in texts based on various probabilistic samples and metrics. And packages such as {maps} or {maptools} or {ggplot2} can be used for geospatial maps, where geographical and quantitative data can be analyzed and overlaid in the same visualization.

R was also successfully used to develop computer simulations such as agent-based modeling and dynamic social network analysis. Some examples are structurally cohesive network blocks (Padgett 2006) or the hypercycles model for economic production as chemistry (Padgett et al. 2003).

R can also be used to do machine learning or Bayesian networks or Bayesian analyses, thus extending the power of the software beyond its original goal of statistical software.

Limitations

Besides the steeper learning curve for beginners, R does not have too many limitations regarding what it can do in terms of data analysis, visualizations, data architectures, data cleaning, and Big Data processing in general. Some limitations may come from packages that are not being updated or maintained and some other limitations are given by memory management, as some tasks or visualizations may take longer computational time to process, making R less than ideal for some data

mining projects. But, in general, R is a very versatile and widely used software for a multitude of analyses and data types.

Further Reading

- Cook, J. D. (2017). R programming for those coming from other languages. Web resource: https://www.johndcook.com/R_language_for_programmers.html. Retrieved 12 May 2017.
- Data Science Wars. https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis#gs.KOD6_nA. Retrieved 12 May 2017.
- Matloff, N. (2011). *The art of R programming: A tour of statistical software design*. New York: No Starch Press.
- Muenchen, R. A. (2012). The popularity of data analysis software. <http://r4stats.com/popularity>.
- Padgett, J. F. (2006). Organizational genesis in Florentine history: Four multiple-network processes (unpublished).

Available at: <https://www.chicagobooth.edu/socialorg/docs/padgett-organizationalgenesis.pdf>.

- Padgett, J. F., Lee, D., & Collier, N. (2003). Economic production as chemistry. *Industrial and Corporate Change*, 12(4), 843–877.
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Raim, A.M. (2013). *Introduction to distributed computing with pbdR at the UMBC High Performance Computing Facility* (PDF). Technical report. UMBC High Performance Computing Facility, University of Maryland, Baltimore. HPCF-2013-2.

Rule

- [Regulation](#)

Salesforce

Jason Schmitt

Communication and Media, Clarkson University,
Potsdam, NY, USA

Salesforce is a global enterprise software company, with Fortune 100 standing, most well-known for its role in linking cloud computing to on-demand customer relationship management (CRM) products. Salesforce CRM and marketing products work together to make corporations more functional and ultimately more efficient. Founded in 1999 by Marc Benioff, Parker Harris, Dave Moellenhoff, and Frank Domingues, Salesforce's varied platforms allow organizations to understand the consumer and the varied media conversations revolving around a business or brand. According to *Forbes* (April 2011) which conducted an assessment of businesses focused on value to shareholders, Marc Benioff of Salesforce was the most effective CEO in the world.

Salesforce provides a cloud-based centralized location to track data. Contacts, accounts, sales deals, and documents as well as corporate messaging and the varied social media conversations are all archived and retrievable within the Salesforce architecture from any web or mobile device without the use of any tangible software. Salesforce's quickly accessible information has

an end goal to optimize profitability, revenue, and customer satisfaction by orientating the organization around the customer. This ability to track and message correctly highlights Salesforce's unique approach to management practice known in software development as Scrum.

Scrum is an incremental software development framework for managing product development by a development team that works as a unit to reach a common goal. A key principle of Salesforce's Scrum direction is the recognition that during a project the customers can change their minds about what they want and need, often called churn, and predictive understanding is hard to accomplish. As such, Salesforce takes an empirical approach in accepting that an organization's problem cannot be fully understood or defined and instead focuses on maximizing the team's ability to deliver messaging quickly and respond to emerging requirements.

Salesforce provides a fully customizable user interface for custom adoption and access for a diverse array of organization employees. Further, Salesforce has the ability to integrate into existing websites and allows for building additional web pages through the cloud-based service. Salesforce has the ability to link with Outlook and other mail clients to sync calendars and associate emails with the proper contact and provides the functionality to keep a record every time a contact or data entry is accessed or amended. Similarly, Salesforce

keeps track and organizes customer support issues and tracks them through to resolution with the ability to escalate individual cases based on time sensitivity and the hierarchy of various clients. Extensive reporting is a value of Salesforce's offerings, which provides management an ability to track problem areas within an organization to a distinct department, area, or tangible product offering.

Salesforce has been a key leader in evolving marketing within this digital era through the use of specific marketing strategy aimed at creating and tracking marketing campaigns as well as measuring the success of online campaigns. These services are part of another growing segment available within Salesforce offerings in addition to the CRM packaging. Marketing departments leveraging Salesforce's Buddy Media, Radian6, or ExactTarget obtain the ability of users to conduct demographic, regional, or national searches on keywords and themes across all social networks, which create a more informed and accurate marketing direction. Further, Salesforce's dashboard, which is the main user interactive page, allows the creation of specific marketing directed tasks that can be customized and shared for differing organizational roles or personal preferences.

Salesforce marketing dashboard utilizes widgets that are custom, reusable page elements, which can be housed on individual users' pages. When a widget is created, it is added to a widgets view where all team members can easily be assigned access. This allows companies and organizations to share appropriate widgets defined and created to serve the target market or industry-specific groups. The shareability of widgets allows the most pertinent and useful tasks to be replicated by many users within a single organization.

Types of Widgets

The Salesforce Marketing Cloud "River of News" is a widget that allows users to scroll through specific search results, within all social media

conversations, and utilizes user-defined keywords. Users have the ability to see original posts that were targeted from keyword searches and provided a source link to the social media platform the post or message originated from. The "River of News" displays posts with many different priorities, such as newest post first, number of Twitter followers, social media platform used, physical location, and Klout score. This tool provides strong functionality for marketers or corporations wishing to hone in, or take part in, industry, customer, or competitor conversations.

"Topic analysis" is a widget that is most often used to show share of voice or the percentage of conversation happening about your brand or organization in relation to competitor brands. It is displayed as a pie chart and can be segmented multiple ways based on user configuration. Many use this feature as a quick visual assessment to see the conversations and interest revolving around specific initiatives or product launches.

"Topic trends" is a widget that provides the ability to display the volume of conversation over time through graphs and charts. This feature can be used to understand macro day, week, or month data. This widget is useful when tracking crisis management or brand sentiment. With a line graph display, users can see spikes of activity and conversation around critical areas. Further, users then can click and hone in on spikes, which can open a "Conversation Cloud" or "River of News" that allows users to see the catalyst behind the spike of social media activity. This tool is used as a way to better understand reasons for increased interest or conversation across broad social media platforms.

Salesforce Uses

Salesforce offers wide ranging data inference from its varied and evolving products. As CRM integration within the web and mobile has increased, the broad interest to better understand and leverage social media marketing campaigns

has risen as well, allowing Salesforce a leading push within this industry's market share. The diverse array of businesses, nonprofits, municipalities, and other organizations that utilize Salesforce illustrates the importance of this software within daily business and marketing strategy. Salesforce clients include the American Red Cross, the City of San Francisco, Philadelphia's 311 system, Burberry, H&R Block, Volvo, and Wiley Publishing.

Salesforce Service Offerings

Salesforce is a leader within other CRM and media marketing-orientated companies such as Oracle, SAP, Microsoft Dynamics CRM, Sage CRM, Goldmine, Zoho, Nimble, Highrise, Insight.ly, and Hootsuite. Salesforce's offerings can be purchased individually or as a complete bundle. It offers current breakdowns of services and access in its varied options that are referred to as *Sales Cloud*, *Service Cloud*, *ExactTarget Marketing Cloud*, *Salesforce1 Platform*, *Chatter*, and Work.com.

Sales Cloud allows businesses to track customer inquiries, escalate issues requiring specialized support, and monitor employee productivity. This product provides customer service teams with the answers to customers' questions and the ability to make the answers available on the web so consumers can find answers for themselves.

Service Cloud offers active and real-time information directed toward customer service. This service provides functionality such as Agent Console which offers relevant information about customers and their media profiles. This service also provides businesses the ability to give customers access to live agent web chats from the web to ensure customers can have access to information without a phone call.

ExactTarget Marketing Cloud focuses on creating closer relationships with customers through directed email campaigns, in-depth social marketing, data analytics, mobile campaigns, and marketing automation.

Sales1Platform is geared toward mobile app creation. Sales1Platform gives access to create and promote mobile apps with over four million apps created utilizing this service.

Chatter is a social and collaborative function that relates to the Salesforce platform. Similar to Facebook and Twitter, Chatter allows users to form a community within their business that can be used for secure collaboration and knowledge sharing.

Work.com is a corporate performance management platform for sales representatives. The platform targets employee engagement in three areas: alignment of team and personal goals with business goals, motivation through public recognition, and real-time performance feedback.

Salesforce has more than 5,500 employees, revenues of approximately \$1.7 billion, and a market value of approximately \$17 billion. The company regularly conducts over 100 million transactions a day and has over 3 million subscribers.

Headquartered in San Francisco, California, Salesforce also maintains regional offices in Dublin, Singapore, and Tokyo with secondary locations in Toronto, New York, London, Sydney, and San Mateo, California. Salesforce operates with over 170,000 companies and 17,000 nonprofit organizations. In June 2004, Salesforce was offered on the New York Stock Exchange under the symbol CRM.

Cross-References

- [Data Aggregation](#)
- [Data Streaming](#)
- [Social Media](#)

Further Reading

Denning, S. (2011). Successfully implementing radical management at Salesforce.com. *Strategy & Leadership*, 39(6), 4.

Satellite Imagery/Remote Sensing

Carolynne Hultquist

Geoinformatics and Earth Observation

Laboratory, Department of Geography and
Institute for CyberScience, The Pennsylvania
State University, University Park, PA, USA

Definition

Remote sensing is a technological approach used to acquire observations of the surface of the Earth and the atmosphere. Remote sensing data is stored in diverse collections on a massive scale, from a variety of platforms and sensors, and at varying spatial and temporal resolutions. The term is often used interchangeably with satellite imagery which uses sensors deployed on satellite platforms to collect observations, but remote sensing imagery can also be collected by manned and unmanned aircraft as well as ground-based sensors. One of the fundamental computational problems in remote sensing is dividing imagery into meaningful groups of features. Methods have developed and been adopted to classify and cluster features in images based on pixel values. In the face of increased imagery resolution and big data, recent approaches involve object-oriented segmentation and machine learning algorithms.

Introduction

The basic principles of remote sensing are related to the sensor itself, the digital products it creates, and the methods used to extract information from this data. The concept of remote sensing is that data are collected without being in contact with the features observed. Typically, imagery is collected from sensors that are on satellite platforms, manned aircraft, and unmanned aerial vehicles (UAVs). The sensors record digital imagery within a grid of pixels that have values from 0 to 255 (Campbell 2011). Remote sensing instruments are calibrated to collected

measurements at different wavelengths in the electromagnetic spectrum which are recorded as bands. Each band records the magnitude of the radiation as the brightness of pixel in the scene. Using a combination of these bands, imagery analysis can be used to identify features on the surface of the Earth. Remote sensing classification techniques are typically used to extract features into classes based on spectral characteristics of imagery of the Earth's surface.

The data mining field of image processing is conceptually similar to remote sensing imagery classification. In image processing, automated processing of visible spectrum RGB (red-green-blue) images use patterns based on identified features in images in order to recognize an overall class to which it belongs. For example, finding all the dogs in a set of images is learned by picking out features of a dog and then identifying dogs by these characteristic features. At even a glance, humans are very good at visual recognition by quickly putting together information perceived to identify objects in images. In these fields of image classifications, individual features are used to identify the overall pattern by moving windows that consider the neighboring pixels in order to pick out spatially close parts of features. Standard image classification may look for the features that make up a face, whereas in remote sensing, the entire landscape is described by classifying the features that make it up based on learned characteristics of those features. For example, we know that vegetation and urban features register as particular spectral signatures in certain bands which allows for the characterization of those features. When those features are extracted with classification, it is based off these learned characteristics of the imagery.

Remote sensing analysis employs recognition of the visual scene; however, it is different from what is normally experienced by humans as the perspective of the image is observed as if above the Earth and characteristics of features not visible to the human eye are made useful by digital observation. Remote sensing classification can take advantage of sensor capabilities to go beyond the visible spectrum by incorporating

available spatial data about the surface of the Earth from bands of infrared, radar, LIDAR (Light Detection and Ranging), etc. Hyperspectral sensors record bands at many small sections of the electromagnetic spectrum. These additional features are not necessary for many applications but can improve classification accuracy as it is more challenging to accurately classify the complexity of features on the Earth's surface with only visible bands.

Historical Background

Remote sensing is a term often used interchangeably with satellite imagery which uses sensors deployed on satellite platforms to collect observations. Yet, remote sensing had its roots in less stable airborne sensors such as pigeons and balloons. As aircraft capabilities advanced over the twentieth century, both manned and unmanned planes were often used for remote sensing applications. Modern satellite programs began to develop for research purposes to observe the environment. For example, the prominent Landsat program developed in the 1970s by the U.S. government to provide satellite remote sensing of the Earth's surface. UAVs are becoming increasingly popular for high-resolution collection as the cost to buy into the technology has decreased and performance capabilities have significantly developed.

Classification of remote sensing imagery to extract features has been developed as a technique since the 1960s. Traditionally, image classification was performed manually by image interpreters who went over imagery section by section to produce useful classes such as land use and land cover. Skilled human interpreters rely on eight elements of imagery interpretation: image tone, texture, shadow, pattern, association, shape, size, and site (Olson 1960). These elements guide human interpreters to define each pixel to a class.

Today, remote sensing imagery is available for download in large quantities online. There is more imagery available and at a higher quality than ever before so automated methods of classification are essential for processing. Automated classification techniques continue to make use of these elements

of imagery interpretation as variables to computationally determine the resulting classification.

Methods

Dividing the image into meaningful groups is a fundamental computational problem in remote sensing analysis. Bands from imagery are stored as pixel values collected by the sensor at a particular wavelength. These spectral values, referred to as digital numbers, can be used to identify features in imagery by classifying the pixels individually or by moving windows that consider the neighboring pixels. Many computational methods were developed over the years to classify and cluster features in images based on pixel values. In light of big data and high spatial resolution, recent approaches involve object-oriented segmentation and machine learning algorithms.

For years, the pixel-based approach was standard and it classified features only at the scale at which features are able to be observed, which were very coarse compared to high spatial resolution modern day sensors. The spatial resolution (measured as the square meter area covered by each pixel) of the imagery from the Landsat satellite system, which has been operational since the 1970s, has traditionally been at 30m by 30m. So, if a building covers less than half of the surface area of 30 by 30m, then it would not be identified as a building. A feature will likely not be classified correctly until it is twice the size of the image resolution as the pixels do not perfectly align with features. An important concept from the discipline of cartography is that coarser grids are used at larger scales with less information content while finer grids are traditionally used only at smaller scales and make details of features identifiable (Hengl 2006). Therefore, Landsat is mostly used for large-scale classification of general land cover categories and not as often used for detailed land use classes that break down types of classes into hierarchies, such as urban classes being made up of impervious surfaces and buildings. Some fuzzy classification methods have been used to improve the pixel-based method.

Modern sensors provide imagery that is at a much higher spatial resolution than previously available. At a spatial resolution of 1.5m

such as SPOT 6, features to be extracted are made up of many pixels. This increase in spatial resolution has caused a major shift in the field from using a pixel-based approach to an object-oriented approach as there are many pixels which make up features. Objects can group together many similar pixels that form a feature, whereas pixel-based approaches leave “edge effects” of misclassified pixels (Campbell 2011). In a pixel-based approach, pixels are individually assigned to a class so small variations in spectral properties can create areas with non-contextual class combinations.

Object-oriented approaches solve this pixel misclassification problem by first segmenting the image pixels into meaningful objects before classification. An object-oriented method segments homogeneous regions of pixels that consider characteristics of at least spatial configuration and spectral characteristics, but often shape, texture, size, and topographical as well. These segmentation parameters are set by the user based on an understanding of the resolution of the imagery being used and the size of the features that are to be identified. Borna et al. (2016) address the issue of subjective scale for object segmentation as the parameters set by trial and error for the size of the objects is shown to affect the resulting classification. In addition, multiple segmentations can be run to construct features that exist at different typical sizes and shapes such as rivers, streams, and lakes.

Machine learning in remote sensing is a growing field and many researchers are turning to automated methods to extract features. Automated classification methods are needed due to the increase in imagery available over large coverage areas at better spatiotemporal resolution. This means that there are more images of high quality (spatial resolution) at more places (coverage) more often (temporal resolution). Using image interpretation techniques from a manual interpreter is a huge time investment and involves human error. Automated methods can provide quick classification with consistent computational accuracy so that we can take advantage of the high spatio-temporal resolution to detect patterns and

changes. The key then becomes accuracy assessment to show the end users that the results of the machine learning techniques are reliable.

Machine learning for image classification traditionally takes advantage of commonly used techniques such as basic Decision Trees to more advanced techniques of Random Forest and boosted trees, Support Vector Machines (SVM), and Gaussian Processes. For hyperspectral imagery, SVM is shown to outperform traditional statistical and nearest neighbor (NN)-based classification with only ensemble methods sometimes having better accuracy when using spatial and textural-based morphological parameters (Chutia et al. 2016). Naive Bayesian classifiers are less used in remote sensing; while the method is fast, they perform poorly as they assume independence between attributes which is not true of imagery.

Feature reduction is necessary in some cases in which many imagery bands are available. Multispectral imagery typically has at the minimum three bands of RGB (red-green-blue). Often other bands are available in the infrared which can be helpful in distinguishing between vegetation and human constructed features. Multispectral imagery is traditionally used for image classification, but recently hyperspectral imagery is becoming available over some areas for many applications that try to observe specific signatures. As Chutia et al. (2016) describes, using hyperspectral imagery creates challenges for classification due to having many bands of imagery; typical systems have 220–400 bands collected at many wavelengths which are often autocorrelated features. Many methods are used for classification feature reduction like decision fusion, mixture modelling, and discriminant analysis (Chutia et al. 2016). The predictive power of methods is increased by reducing the high dimensionality of the hyperspectral bands and having low linear correlation between bands using techniques such as principle components (Chutia et al. 2016), smart band selection (SBS), and minimum redundancy maximum relevance (mRMR). Instead of reducing the features for traditional classification techniques, other

methods can be used such as deep learning which takes advantage of the high dimensionality of data available.

Applications

Remote sensing is used for a variety of applications in order to identify or measure features and detect changes. These classes to be identified can be land use/land cover types, binary detection of change, levels of damage assessment, specific mineral types, etc. Some application areas are measuring the extent of impact from disasters, the melting of glaciers, mapping geology, or urban land use planning. There is a growing field of interest in using remote sensing to analyze and optimize agriculture. Archeological applications are using technologies such as UAV's, photogrammetry for 3D modeling, and radar for buried sites. Environmental change can be monitored remotely over large areas. Meteorological conditions are constantly monitored with satellite imagery. As the technology advances, many new application fields are developing.

Conclusion

Remote sensing is a growing technological field with methodological advancements to meet the computational need for processing big data. The use of remote sensing for specialized applications is becoming publically accessible with growing interest in how to use the data and decreasing costs to buy into the technology. Hopefully the field will continue to develop to meet needs in critical application areas. Remote sensing draws users in as it enables us to look beyond what we can naturally see to identify features of interest and recognize measureable changes in the environment.

Cross-References

- [Environment](#)
- [Sensor Technologies](#)

Further Reading

- Borna, K., Moore, A. B., & Sirguey, P. (2016). An intelligent geospatial processing unit for image classification based on geographic vector agents (GVAs). *Transactions in GIS*, 20(3), 368–381. <http://doi.org/10.1111/tgis.12226>.
- Campbell, J. B. (2011). *Introduction to remote sensing* (5th ed.). New York: The Guilford Press. ISBN 978-1609181765.
- Chutia, D., Bhattacharyya, D. K., Sarma, K. K., Kalita, R., & Sudhakar, S. (2016). Hyperspectral remote sensing classifications: A perspective survey. *Transactions in GIS*, 20(4), 463–490. <http://doi.org/10.1111/tgis.12164>.
- Hengl, T. (2006). Finding the right pixel size. *Computers & Geosciences*, 32(9), 1283–1298.
- Olson, C. E. (1960). Elements of photographic interpretation common to several sensors. *Photogrammetric Engineering*, 26(4), 651–656.

Scientometrics

Jon Schmid
Georgia Institute of Technology, Atlanta,
GA, USA

Scientometrics refers to the study of science through the measurement and analysis of researchers' productive outputs. These outputs include journal articles, citations, books, patents, data, and conference proceedings. The impact of big data analytics on the field of scientometrics has primarily been driven by two factors: the emergence of large online bibliographic databases and a recent push to broaden the evaluation of research impact beyond citation-based measures. Large online databases of articles, conferences proceedings, and books allow researchers to study the manner in which scholarship develops and measure the impact of researchers, institutions, and even countries on a field of scientific knowledge. Using data on social media activity, article views, downloads, social bookmarking, and the text posted on blogs and other websites, researchers are attempting to broaden the manner in which scientific output is measured.

Bibliometrics, a subdiscipline of scientometrics that focuses specifically on the study of scientific publications, witnessed a boon in research due to the emergence of large digital bibliographic databases such as Web of Science, Scopus, Google Scholar, and PubMed. The utility of increased digital indexing is enhanced by the recent surge in total scientific output. Lutz Bornmann and Ruediger Mutz find that global scientific output has grown at a rate of 8–9% per year since World War II (equivalent to a doubling every 9 years) (Bornmann and Mutz 2015).

Bibliometric analysis using large data sets has been particularly useful in research that seeks to understand the nature of research collaboration. Because large bibliographic databases contain information on coauthorships, the institutions that host authors, journals, and publication dates, text mining software can be used in combination with social network analysis to understand the nature of collaborative networks. Visualizations of these networks are increasingly used to show patterns of collaboration, ties between scientific disciplines, and the impact of scientific ideas. For example, Hanjun Xian and Krishna Madhavan analyzed over 24,000 journal articles and conference proceedings from the field of engineering education in effort to understand how the literature was produced (Xian and Madhavan 2014). These data were used to map the network of collaborative ties in the discipline. The study found that cross-disciplinary scholars played a critical role in linking isolated network segments.

Besides studying authorship and collaboration, big data analytics have been used to analyze citations to measure the impact of research, researchers, and research institutions. Citations are a common proxy for the quality of research. Important papers will generally be highly cited as subsequent research relies on them to advance knowledge.

One prominent metric used in scientometrics is the *h*-index, which was proposed by Jorge Hirsch in 2005. The *h*-index considers the number of publications produced by an individual or organization and the number of citations these publications receive. An individual can be said to have an *h*-index of *h* when she produces *h* publications

each of which receives at least *h* citations and no other publication receives more than *h* citations.

The advent of large databases and big data analytics has greatly facilitated the calculation of the *h*-index and similar impact metrics. For example, in a 2013 study, Filippo Radicchi and Claudio Castellano utilized the Google Scholar Citations data set to evaluate the individual scholarly contribution of over 35,000 scholars (Radicchi and Castellano 2013). The researchers found that the number of citations received by a scientist is a strong proxy for that scientist's *h*-index, whereas the number of publications is a less precise proxy.

The same principles behind citation analysis can be applied to measure the impact or quality of patents. Large patent databases such as PATSTAT allow researchers to measure the importance of individual patents using forward citations. Forward citations come from the “prior art” section of the patent documents, which describes the technologies that were deemed critical to their innovation by the patent applicants. Scholars use patent counts, weighed by forward citations, to derive measures of national innovative productivity.

Until recently, measurement of research impact has been almost exclusively based on citation-based measures. However, citations are slow to accumulate and ignore the influence of research on the broader public. Recently there has been a push to include novel data sources in the evaluation of research impact. Gunther Eysenbach has found that tweets about a journal article within the first 3 days of publication are a strong predictor of eventual citations for highly cited research articles (Eysenbach 2011). The direction of causality in this relationship – i.e., whether strong papers lead to a high volume of tweets or whether the tweets themselves cause subsequent citations – is unclear. However, the author suggests that the most promising use of social media data lies not in its use as a predictor of traditional impact measures but as means of creating novel metrics of the social impact of research.

Indeed the development of an alternative set of measurements – often referred to as “altmetrics” – based on data gleaned from the social web represents a particularly active field of scientometrics research. Toward this end, services such as PLOS Article-Level Metrics use big data techniques to

develop metrics of research impact that consider factors other than citations. PLOS Article-Level Metrics pulls in data on article downloads, commenting and sharing via services such as CiteULike, Connotea, and Facebook, to broaden the way in which a scholar's contribution is measured.

Certain academic fields, such as the humanities, that rely on under-indexed forms of scholarship such as book chapters and monographs have proven difficult to study using traditional scientometrics techniques. Because they do not depend on online bibliographic databases, altmetrics may prove useful in studying such fields. Björn Hammarfelt uses data from Twitter and Mendeley – a web-based citation manager that has a social networking component – to study scholarship in the humanities (Hammarfelt 2014). While his study suggests that coverage gaps still exist using altmetrics, as these applications become more widely used, they will likely become a useful means of studying neglected scientific fields.

Cross-References

- [Bibliometrics/Scientometrics](#)
- [Social Media](#)

Further Reading

- Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215–2222. arXiv:1402.4578 [Physics, Stat].
- Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13, e123.
- Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics*, 101, 1419–1430.
- Radicchi, F., & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, 97(3), 627–637. <https://doi.org/10.1007/s11192-013-1027-3>.
- Xian, H., & Madhavan, K. (2014). Anatomy of scholarly collaboration in engineering education: A big-data bibliometric analysis. *Journal of Engineering Education*, 103, 486–514.

Semantic Data Model

- [Ontologies](#)

Semantic/Content Analysis/ Natural Language Processing

Paul Nulty

Centre for Research in Arts Social Science and Humanities, University of Cambridge, Cambridge, United Kingdom

Introduction

One of the most difficult aspects of working with big data is the prevalence of unstructured data, and perhaps the most widespread source of unstructured data is the information contained in text files in the form of natural language. Human language is in fact highly structured, but although major advances have been made in automated methods for symbolic processing and parsing of language, full computational language understanding has yet to be achieved, and so a combination of symbolic and statistical approaches to machine understanding of language are commonly used. Extracting meaning or achieving understanding from human language through statistical or computational processing is one of the most fundamental and challenging problems of artificial intelligence. From a practical point of view, the dramatic increase in availability of text in electronic form means that reliable automated analysis of natural language is an extremely useful source of data for many disciplines.

Big data is an interdisciplinary field, of which natural language processing (NLP) is a fragmented and interdisciplinary subfield. Broadly speaking, researchers use approaches somewhere on a continuum between representing and parsing the structures of human language in a symbolic, rule-based fashion, or feeding large amounts of minimally preprocessed text into more sophisticated statistical machine learning

systems. In addition, various substantive research areas have developed overlapping but distinct methods for computational analysis of text.

The question of whether NLP tasks are best approached with statistical, data-driven methods or symbolic, theory-driven models is an old debate. In 1957, Noam Chomsky wrote:

it must be recognized that the notion of “probability of a sentence” is an entirely useless one, under any known interpretation of this term.

However, at present the best methods we have for translating, searching, and classifying natural language text use flexible machine-learning algorithms that learn parameters probabilistically from relatively unprocessed text. On the other hand, some applications, such as the IBM Watson question answering system (Ferruci et al. 2010), make good use of a combination of probabilistic learning and modules informed by linguistic theory to disambiguate nuanced queries.

The field of computational linguistics originally had the goal of improving understanding of human language using computational methods. Historically, this meant implementing rules and structures inspired by the cognitive structures proposed by Chomskyan generative linguistics. Over time, computational linguistics has broadened to include diverse methods for machine processing of language irrespective of whether the computational models are plausible cognitive models of human language processing. As practiced today, computational linguistics is closer to a branch of computer science than a branch of linguistics. The branch of linguistics that uses quantitative analysis of large text corpora is known as corpus linguistics.

Research in computational linguistics and natural language processing involves finding solutions for the many subproblems associated with understanding language, and combining advances in these modules to improve performance on general tasks. Some of the most important NLP subproblems include part-of-speech tagging, syntactic parsing, identifying the semantic roles played by verb arguments, recognizing named entities, and resolving references. These feed into performance on more general tasks like

machine translation, question answering, and summarization.

In the social sciences, the terms quantitative content analysis, quantitative text analysis, or “text as data” are all used. Content analysis may be performed by human coders, who read and mark-up documents. This process can be streamlined with software. Fully automated content analysis, or quantitative text analysis, typically employs statistical word-frequency analysis to discover latent traits from text, or scale documents of interest on a particular dimension of interest in social science or political science.

Tools and Resources

Text data does not immediately challenge computational resources to the same extent as other big data sources such as video or sensor data. For example, the entire proceedings of the European parliament from 1996 to 2005, in 21 languages, can be stored in 5.4 gigabytes – enough to load into main memory on most modern machines. While techniques such as parallel and distributed processing may be necessary in some cases, for example, global streams of social media text or applying machine learning techniques for classification, typically the challenge of text data is to parse and extract useful information from the idiosyncratic and opaque structures of natural language, rather than overcoming computational difficulties simply to store and manipulate the text. The unpredictable structure of text files means that general purpose programming languages are commonly used, unlike in other applications where the tabular format of the data allows the use of specialized statistical software.

Original Unix command line tools such as `grep`, `sed`, and `awk` are still extremely useful for batch processing of text documents. Historically, Perl has been the programming language of choice for text processing, but recently Ruby and Python have become more widely used. These are scripting languages, designed for ease of use and flexibility rather than speed. For more computationally intensive tasks, NLP tools are implemented in Java or C/C++.

The python libraries spaCy and gensim and the Java-based Stanford Core NLP software are widely used in industry and academia. They provide implementations and guides for the most widely used text processing and statistical document analysis methods.

Preprocessing

The first step in approaching a text analysis dataset is to successfully read the document formats and file encodings used. Most programming languages provide libraries for interfacing with Microsoft Word and pdf documents. The ASCII coding system represents unaccented English upper and lowercase letters, numbers, and punctuation, using one byte per character. This is no longer sufficient for most purposes, and modern documents are encoded in a diverse set of character encodings. The Unicode system defines code points which can represent characters and symbols from all writing systems. The UTF-8 and UTF-16 encodings implement these code points in 8 bit or 16 bit encoded files.

Words are the most apparent units of written text, and most text processing tasks begin with tokenization – dividing the text into words. In many languages, this is relatively uncomplicated: whitespace delimits words, with a few ambiguous cases such as hyphenation, contraction, and the possessive marker. Within languages written in the Roman alphabet there is some variance, for example, agglutinative languages like Finnish and Hungarian tend to use long compound terms disambiguated by case markers, which can make the connection between space-separated words and dictionary-entry meanings tenuous. For languages with a different orthographic system, such as Chinese, Japanese, and Arabic, it is necessary to use a customized tokenizer to split text into units suitable for quantitative analysis.

Even in English, the correspondence between space-separated word and semantic unit is not exact. The fundamental unit of vocabulary – sometimes called the lexeme – may be modified or inflected by the addition of morphemes indicating tense, gender, or number. For many

applications, it is not desirable to distinguish between the inflected forms of words, rather we want to sum together counts of equivalent words. Therefore, it is common to remove the inflected endings of words and count only the root, or stem. For example, a system to judge the sentiment of a movie review need not distinguish between the words “excite,” “exciting,” “excites,” and “excited.” Typically the word ending is removed and the terms are treated equivalently.

The Porter stemmer (Porter 1980) is one of the most frequently used algorithms for this purpose. A slightly more sophisticated method is lemmatization, which also normalizes inflected words, but uses a dictionary to match irregular forms such as “be”/“is”/“are”. In addition to stemming and tokenizing, it may be useful to remove very common words that are unlikely to have semantic content related to the task. In English, the most common words are function words such as “of,” “in,” and “the.” These “stopwords” largely serve a grammatical rather than semantic function, and some NLP systems simply remove them before proceeding with a statistical analysis.

After the initial text preprocessing, there are several simple metrics that may be used to assess the complexity of language used in the documents. The type-token ratio, a measure of lexical diversity, gives an estimate of the complexity of the document by comparing the total number of words in the document to the number of unique words (i.e., the size of the vocabulary). The Fleisch-Kincaid readability metric uses the average sentence length and the average number of syllables per word combined with coefficients calibrated with data from students to give an estimate of the grade-level reading difficulty of a text.

Document-Term Matrices

After tokenization and other preprocessing steps, most text analysis methods work with a matrix that stores the frequency with which each word in the vocabulary occurs in each document. This is the simplest case, known as the “bag-of-words” model, and no information about the ordering of the words in the original texts is retained. More

sophisticated analysis might involve extracting counts of complex features from the documents. For example, the text may be parsed and tagged with part-of-speech information as part of the preprocessing stage, which would allow for the words with identical spellings but different part-of-speech categories or grammatical roles to be counted as separate features.

Often, rather than using only single words, counts of phrases are used. These are known as n -grams, where n is the number of words in the phrase, for example, trigrams are three-word sequences. N -gram models are especially important for language modeling, used to predict the probability of a word or phrase given the preceding sequence of words. Language modeling is particularly important for natural language generation and speech recognition problems.

Once each document has been converted to a row of counts of terms or features, a wide range of automated document analysis methods can be employed. The document-term matrix is usually sparse and uneven – a small number of words occur very frequently in many documents, while a large number of words occur rarely, and most words do not occur at all in a given document. Therefore, it is common practice to smooth or weight the matrix, either using the log of the term frequency or with a measure of term importance like *tf-idf* (term frequency $\times$ inverse document frequency) or mutual information.

Matrix Analysis

Supervised classification methods attempt to automatically categorize documents based on the document-term matrix. One of the most familiar of such tasks is the email spam detection problem. Based on the frequencies of words in a corpus of emails, the system must decide if an email is spam or not. Such a system is supervised in the sense that it requires as a starting point a set of documents that have been correctly labeled with the appropriate category, in order to build a statistical model of which terms are associated with each category. One simple and effective algorithm for

supervised document classification is Naive Bayes, which gives a new document the class that has the maximum a posteriori probability given the term counts and their independent association between the terms and the categories in the training documents. In political science, a similar algorithm – “wordscores” – is widely used, which sums Naive-Bayes-like word parameters to scale new documents based on reference scores assigned to training texts with extreme positions (Laver et al. 2003).

Other widely used supervised classifiers include support vector machines, logistic regression, and nearest neighbor models. If the task is to predict a continuous variable rather than a class label, then a regression model may be used. Statistical learning and prediction systems that operate on text data very often face the typical big data problem of having more features (word types) than observed or labeled documents. This is a high dimensional learning problem, where p (the number of parameters) is much larger than n (the number of observed examples).

In addition, word frequencies are extremely unevenly distributed (an observation known as Zipf’s law) and are highly correlated with one another, resulting in parameter vectors that make less than ideal examples for regression models. It may therefore be necessary to use regression methods designed to mitigate this problem, such as lasso and ridge regression, or to prune the feature space to avoid overtraining, using feature subset selection or a dimensionality reduction technique like principal components analysis or singular value decomposition. With recent advances in neural network research, it has become more common to use unprocessed counts of n -grams, tokens, or even characters as input to a neural network with many intermediate layers. With sufficient training data, such a network can learn the feature extraction process better than hand-curated feature extraction systems, and these “deep learning” networks have improved the state of the art in machine translation and image labeling.

Unsupervised methods can cluster documents or reveal the distribution of topics in documents in a data-driven fashion. For unsupervised scaling

and clustering of documents, methods include k-means clustering, or the Wordfish algorithm, a multinomial Poisson scaling model for political documents.

Another goal of unsupervised analysis is to measure what topics comprise the text corpus, and how these topics are distributed across documents. Topic modeling (Blei 2012) is a widely used generative technique to discover a set of topics that influence the generation of the texts, and explore how they are associated with other variables of interest.

Vector Space Semantics and Machine Learning

In addition to retrieving or labeling documents, it can be useful to represent the meaning of terms found in the documents. Vector space semantics, or distributional semantics, aims to represent the meaning of words using counts of their co-occurrences with other words. The “distributional hypothesis,” as described by JR Firth (Firth 1957), is the idea that “you shall know a word by the company it keeps.” The co-occurrence vectors of words have been shown to be useful for noun phrase disambiguation, semantic relation extraction, and analogy resolution. Many systems now use the factorization of the co-occurrence matrices as the initial input to statistical learners, allowing a fine-grained representation of lexical semantics. Vector semantics also allows for word sense disambiguation – it is possible to distinguish the different senses of a word by clustering the vector representations of its occurrences.

These vectors may count instances of words co-occurring with the same context (syntagmatic relations) or compare the similarity of the contexts of words as a measure of their substitutability (paradigmatic relations) (Turney and Pantel 2010). The use of neural networks or dimensionality reduction techniques allows researchers to produce a relatively low dimensional space in which to compare word vectors, sometimes called word embeddings.

Machine learning has long been used to perform classification of documents or to aid the

accuracy of NLP subtasks described above. However, as in many other fields, the recent application of neural networks with many hidden layers (Deep Learning) has led to large improvements in accuracy rates on many tasks. These opaque but computationally powerful techniques require only a large volume of training data and a differentiable target function to model complex linguistic behavior.

Conclusion

Natural language processing is a complex and varied problem that lies at the heart of artificial intelligence. The combination of statistical and symbolic methods has led to huge leaps forward over the last few decades, and with the preponderance of online training data and advances in machine learning methods, it is likely that further gains will be made in the coming years. For researchers intending to make use of rather than advance these methods, a fruitful approach is a good working knowledge of a general purpose programming language, combined with the ability to configure and execute off-the-shelf machine learning packages.

Cross-References

- [Artificial Intelligence](#)
- [Machine Learning](#)
- [Unstructured Data](#)

References

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- Chomsky, N. (2002). *Syntactic structures*. Berlin: Walter de Gruyter.
- Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J., Nyberg, E., Prager, J., Schlaefel, N., & Welty, C. A. (2010). Building Watson: An overview of the deep QA project. *AI Magazine*, 31(3), 59–79.
- Firth, J. R. (1957). A synopsis of linguistic theory. In *Studies in linguistic analysis*. Blackwell: Oxford.

- Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02), 311–331.
- Porter, MF. "An algorithm for suffix stripping." *Program* 14.3 (1980): 130–137.
- Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3), 705–722.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141–188.

Semiotics

Erik W. Kuiler

George Mason University, Arlington, VA, USA

Background

Semiotics, as an intellectual discipline, focuses on the relationships between signs, the objects to which they refer, and the interpreters – human individuals or information and communications technology (ICT) intelligent agents – who assign meaning to the conceptualizations as well as instantiations of such signs and objects based on those relationships. By focusing on diverse kinds of communications as well as their users, semiotics supports the development of a conceptual framework to explore various aspects of knowledge transformation and dissemination that include the reception and manipulation of signs, symbols, and signals received from diverse sources, such as signals received from medical Internet of Things (IoT) devices, and the algorithms and analytics required to transform signals and data into information and knowledge. Semiotics encompasses different aspects of knowledge formulation – the epistemically constrained process of extrapolating signals from noise, transforming signals into data, and, subsequently, into knowledge that can be operationalized using large datasets to support strategic planning, decision-making, and data analytics.

From the perspective of an ICT-focused semiotics, a *signal* is an anomaly discovered in the context of a perceived indiscriminate,

undifferentiated field of noise and is, in effect, recognition of a pattern out of noise. Frequently, signals provide an impetus to action. For example, in an IoT network, signals reflect properties of frequency, duration, and strength that indicate a change in an object's state or environment to elicit a response from an interpreting agent (be it an ICT or human agent) or to transfer meaning.

Signals become *data* – the creation of facts (something given or admitted, especially as a basis for reasoning or inference) – by imposing syntactic and symbolic norms on signals. Data become signs when semantics are applied. *Signs* reflect cultural, epistemic norms and function as morphemes of meaning by representing objects in the mind of the interpreter. In Aristotelian terms, signs and their associated symbols are figures of thought that allow an individual to think about an object without its immediate presence. In this sense, signs have multiple aspects: a designative aspect, which points an interpreter to a specific object (an index); an appraisive aspect, which draws attention to the object's ontological properties; and a prescriptive aspect, which instructs the interpreter to respond in a specific way, such as in response to a transmission stop signal. A *symbol* is a mark, established by convention, to represent an object state, process state, or situation. For example, a flashing red light is usually used to indicate a state of danger or failure. *Information* comprises at least one datum. Assuming that the data are syntactically congruent, information constitutes the transformation of data by attaching meaning to the collection in which the individual data are grouped, as the result of analyzing, calculating, or otherwise exploring them (e.g., by aggregation, combination, decomposition, transformation, correlation, mapping, etc.), usually for assessing, calculating, or planning a course of action. *Knowledge* constitutes the addition of purpose and conation to the understanding gained from analyzing information.

Semiotics: Overview

Semiotics comprises three interrelated disciplines: semantics, syntagmatics (including syntactics),

and pragmatics. Semantics focuses on the sign-object relationships, i.e., the signification of signs and the perception of meaning, for example, by implication, logic, or reference. Syntagmatics focuses on sign-to-sign relationships, i.e., the manner in which signs may be combined to form well-formed composite signs (e.g., well-formed predicates). Pragmatics focuses on sign-interpreter relationships, i.e., methods by which meaning is derived from a sign or combination of signs in a specific context.

Semantics: Lexica and Ontologies

Lexica and ontologies provide the semantics component for a semiotics-focused approach to information derivation and knowledge formulation. Lexica and ontologies reflect social constructions of reality, defined in the context of specific epistemic cultures as sets of norms, symbols, human interactions, and processes that collectively facilitate the transformation of data into information and knowledge. A lexicon functions as a controlled vocabulary and contains the terms and their definitions that collectively constitute the epistemic domain. The terms and their definitions that constitute the lexicon provide the basis for the ontology, which delineates the interdependencies among categories and their properties, usually in the form of similes, meronymies, and metonymies.

Ontologies define and represent the concepts that inform epistemic domains, their properties, and their interdependencies. An ontology, when populated with valid data, provides a base for knowledge formulation that supports the analytics of those data that collectively operationalize that domain. An *ontology* informs a named perspective defined over a set of categories (or classes) that collectively delimit a domain of knowledge. In this context, a *category* delineates a named perspective defined over a set of properties. A *property* constitutes an attribute or characteristic common to these instances that constitute a category; for example, length, diameter, and mode of ingestion. A *taxonomy* is a directed acyclic perspective defined over a set of categories; for example, a hierarchical tree structure depicting the various superordinate,

ordinate, and subordinate categories of an ontology.

Ontologies provide the semantic congruity, consistency, and clarity to support different algorithmic-based aggregations, correlations, and regressions. From an ICT perspective, ontologies enable the development of interoperable information systems.

Syntagmatics: Relationships and Rules

As morphemes of meaning, signs participate in complex relationships. In paradigmatic relations, signs obtain their meaning from their association with other signs based on substitution so that other signs (terms and objects) may be substituted for signs in the predicate, provided that the signs belong to the same ontological category (i.e., paradigmatic relations support lexical alternatives and semantic likenesses). Indeed, the notion of paradigmatic relations is foundational for the development of ontologies by providing the means to develop categories based on properties shared by individual instances.

Syntactics: Metadata

Whereas the lexicon and ontology support the semantic and interpretive aspects of data analytics, metadata support the semantic and syntagmatic operational aspects of data analytics. Metadata are generally considered to be information about data and are usually formulated and managed to comply with predetermined standards. *Operational metadata* reflect the management requirements for data security and safeguarding personally identifiable information; data ingestion, federation, and integration; data anonymization; data distribution; and analytical data storage. *Structural (syntactic) metadata* provide information about data structures (e.g., file layouts or database table and column specifications). *Bibliographical metadata* provide information about the dataset's producer, such as the author, title, table of contents, and applicable keywords of a document; data lineage metadata provide information about the chain of custody of a data item with respect to its provenance – the chronology of data ownership, stewardship, and transformations.

Pragmatics

From the perspective of ICT, pragmatics has two complementary, closely linked components: (1) operationalization support and (2) analytics support. Operationalization pragmatics focuses on the development, management, and governance of ontologies and lexica, metadata, interoperability, etc. Analytical pragmatics focuses on how meaning is derived by engaging rhetoric, hermeneutics, logic, and heuristics and their attendant methods to discern meaning in data, create information, and develop knowledge.

Summary

Semiotics is normative, bounded by epistemic and cultural contexts, and provides the foundation for lexicon and ontology development. Knowledge formulation paradigms depend on lexica and ontologies to provide repositories for formal specifications of the meanings of symbols delineated by the application of semiotics.

Further Reading

- Morris, C. W. (1938). *The foundations of the theory of signs*. Chicago: Chicago University Press.
- Morris, C. W. (1946). Signs, language, and behavior. In C. W. Morris (Ed.), *Writings on the general theory of signs* (pp. 73–398). The Hague: Mouton.
- Morris, C. W. (1964). *Signification and significance: A study of the relations of signs and values*. Cambridge, MA: MIT Press.
- Nöth, W. (1995). *Handbook of semiotics*. Bloomington: Indiana University Press.
- Ogde, C. K., & Richards, I. A. (2013). *The meaning of meaning: A study of language upon thought and the science of symbolism*. CT: Mansfield Centre.
- Peirce, C. S. (1958). *Collected papers of C.S. Peirce*. Cambridge, MA: Harvard University Press.
- Sowa, J. F. (2000). Ontology, metadata, and semiotics. In B. Ganter & G. Mineau (Eds.), *Conceptual structures: Logical, linguistics, and computational issues* (pp. 55–81). Berlin: Springer Verlag.
- Sowa, J. F. (2000). *Knowledge representation: Logical, philosophical, and computational foundations*. Pacific Grove: Brooks/Cole.

Semi-structured Data

Yulia A. Strekalova¹ and Mustapha Bouakkaz²

¹College of Journalism and Communications, University of Florida, Gainesville, FL, USA

²University Amar Telidji Laghouat, Laghouat, Algeria

More and more data become available electronically every day, and they may be stored in a variety of data systems. Some data entries may reside in unstructured document file systems, and some data may be collected and stored in highly structured relational databases. The data itself may represent raw images and sounds or come with a rigid structure as strictly entered entities. However, a lot of data currently available through public and proprietary data systems is semi-structured.

Definition

Semi-structured data is data that resembles structured data by its format but is not organized with the same restrictive rules. This flexibility allows collecting data even if some data points are missing or contain information that is not easily translated in a relational database format. Semi-structured data carries the richness of human information exchange, but most of it cannot be automatically processed and used. Developments in markup languages and software applications allow the collection and evaluation of semi-structured data, but the richness of natural text contained in semi-structured data still presents challenges for analysts.

Structured data has been organized into a format that makes it easier to access and process such as databases where data is stored in columns, which represent the attribute of the database. In reality, very little data is completely structured. Conversely, unstructured data has been not reformatted, and its elements are not organized into a data structure. Semi-structured data combines some elements of both data types. It is not

organized in a complex manner that supports immediate analyses; however, it may have information associated with it, such as metadata tagging, that allows elements contained to be addressed through more sophisticated access queries. For example, a word document is generally considered to be unstructured data. However, when metadata tags in the form of keywords that represent the document content are added, the data becomes semi-structured.

Data Analysis

The volume and unpredictable structure of the available data present challenges in analysis. To get meaningful insights from semi-structured data, analysts need to pre-analyze it to ask questions that can be answered with the data. The fact that a large number of correlations can be found does not necessarily mean that analysis is reliable and complete. One of the preparation measures before the actual data analysis is data reduction. While a large number of data points may be available for collection, not all these data points should be included in an analysis to every question. Instead, a careful consideration of data points is likely to produce a more reliable and explainable interpretation of observed data. In other words, just because the data is available, it does not mean it needs to be included in the analysis. Some elements may be random and will not add substantively to the answer to a particular questions. Some other elements may be redundant and not add any new information compared to the one already provided by other data points.

Jules Berman suggests nine steps to the analysis of semi-structured data. Step 1 includes formulation of a question which can and will be subsequently answered with data. A Big Data approach may not be the best strategy for questions that can be answered with other traditional research methods. Step 2 evaluates data resources available for collection. Data repositories may have “blind spots” or data points that are systematically excluded or restricted for public access. At step 3, a question is reformulated to adjust for the resources identified in step 2. Available data

may be insufficient to answer the original question despite the access to large amounts of data. Step 4 involved evaluation of possible query outputs. Data mining may return a large number of data points, but these data points most frequently need to be filtered to focus on the analysis of the question at hand. At step 5, data should be reviewed and evaluated for its structure and characteristics. Returned data may be quantitative or qualitative, or it may have data points which are missing for a substantial number of records, which will impact future data analysis. Step 6 requires a strategic and systematic data reduction. Although it may sound counterintuitive, Big Data analysis can provide most powerful insights when the data set is condensed to bare essentials to answer a focused question. Some collected data may be irrelevant or redundant to the problem at hand and will not be needed for the analysis. Step 7 calls for the identification of analytic algorithms, should they be deemed necessary. Algorithms are analytic approaches to data, which may be very sophisticated. However, establishing a reliable set of meaningful metrics to answer a question may be a more reliable strategy. Step 8 looks at the results and conclusions of the analysis and calls for conservative assessment of possible explanations and models suggested by the data, assertions for causality, and possible biases. Finally, step 9 calls for validation of results in step 8 using comparable data sets. Invalidation of predictions may suggest necessary adjustments to any of the steps in the data analysis and make conclusions more robust.

Data Management

Semi-structured data includes both database characteristics and incorporates documents and other files types, which cannot be fully described by a standard database entry. Data entries in structured data sets follow the same order; all entries in a group have the same descriptions, defined format, and predefined length. In contrast, semi-structured data entries are organized in semantic entities, similar to the structured data, which may not have same attributes in the same order or of the same length. Early digital databases were

organized based on the relational model of data, where data is recorded into one or more tables with a unique identifier for each entry. The data for such databases needs to be structured uniformly for each record. Semi-structured data but relies on tag or other markers to separate data elements. Semi-structured data may miss data elements or have more than one data point in an element. Overall, while semi-structured data has a predefined structure, the data within this structure is not entered with the same rigor as in the traditional relational databases. This data management situation arises from the practical necessity to handle user-generated and widely interactional data brought up by the Web 2.0. The data contained in emails, blog posts, PowerPoint presentation files, images, and videos may have very different sets of attributes, but they also offer a possibility to assign metadata systematically. Metadata may include information about author and time and may create the structure to assign the data to semantic groups. Unstructured data, on the other hand, is the data that cannot be readily organized in tables to capture the full extent of it. Semi-structured data, as the name suggests, carries some elements of structured data. These elements are metadata tags that may list the author or sender, entry creation and modification times, the length of a document, or the number of slides in a presentation. Yet, these data also have elements that cannot be described in a traditional relational database. For example, traditional database structure which would require initial infrastructure design will not be able to handle information as a sent email, and all response that were received as it is unknown if an email respondents will use one or all names in response, if anyone will get added or omitted, if original message will be modified, if attachments will be added to subsequent messages, etc.

Semi-structured data allows programmers to nest data or create hierarchies that represent complex data models and relationships among entries. However, robustness of the traditional relational data model forces more thoughtful implementation of data applications and possible subsequent ease in analysis. Handling of semi-structured data

is associated with some challenges. The data itself may present a problem by being embedded in natural text, which cannot always be extracted automatically with precision. Natural text is based on sentences which may not have easily identifiable relationships and entities which are necessary for data collection. Natural text is based on sentences that may not have easily identifiable relationships and entities, which are necessary for data collection, and the lack of widely accepted standards for vocabularies. A communication process may involve different models to transfer the same information or require richer data transfer available through natural text and not through a structured exchange of keywords. For example, email exchange can capture the data about senders and recipients, but automated filtering and analysis of the body of email are limited.

Two main types of semi-structured data formats are Extensible Markup Language (XML) and JavaScript Object Notation (JSON). XML, developed in the mid-1990s, is a markup language that sets rules for the data interchange. XML, although being an improvement to earlier markup languages, has been critiqued for being bulky and cumbersome in implementation. JSON is viewed as a possible successor format for digital architecture and database technologies. JSON is an open standard format that transmits data between an application and a server. Data objects in JSON format consist of attribute-value pairs stored in databases like MongoDB and Couchbase. The data, which is stored in a database like MongoDB, can be pulled with a software network for more efficient and faster processing. Apache Hadoop is an example of an open-source framework that provides both storage and processing support. Other multi-platform query processing applications suitable for enterprise-level use are Apache Spark and Presto.

Cross-References

- [Data Integration](#)
- [Digital Storytelling, Big Data Storytelling](#)
- [Discovery Analytics, Discovery Informatics](#)

Further Reading

- Abiteboul, S., et al. (2012). *Web data management*. New York: Cambridge University Press.
- Foreman, J. W. (2013). *Data smart: Using data science to transform information into insight*. Indianapolis: Wiley.
- Miner, G., et al. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham: Academic.

Sensor Technologies

Carolynne Hultquist

Geoinformatics and Earth Observation

Laboratory, Department of Geography and
Institute for CyberScience, The Pennsylvania
State University, University Park, PA, USA

Definition/Introduction

Sensors technologies are developed to detect specific phenomena, behavior, or actions. The origin of the word sensor comes from the Latin root “sentire” a verb defined as “to perceive” (Kalantar-zadeh 2013). Sensors are designed to identify certain phenomena as a signal but not record anything else as it would create noise in the data. Sensors are specified by purpose to identify or measure the presence or intensity of different types of energy: mechanical, gravitational, thermal, electromagnetic, chemical, and nuclear. Sensors have become part of everyday life and continue to grow in importance in modern applications.

Prevalence of Sensors

Sensors are used in everyday life to detect phenomena, behavior, or actions such as force, temperature, pressure, flow, etc. The type of sensor utilized is based on the type of energy that is being sensed, be it gravitational, mechanical, thermal, electromagnetic, chemical, or nuclear. The activity of interest is typically measured by a sensor and converted by a transducer into a signal as a quantity (McGrath

and Scanail 2013). Sensors have been integrated into daily life so that we use them without considering tactile sensors such as elevator buttons, touchscreen devices, and touch sensing lamps. Typical vehicles contain numerous sensors for driving functions, safety, and the comfort of the passengers. Mechanical sensors measure motion, velocity, acceleration, and displacement through such sensors as strain gauges, pressure, force, ultrasonic, acoustic wave, flow, displacement, accelerometers, and gyroscopes (McGrath and Scanail 2013). Chemical and thermal biometric sensors are often used for healthcare from traditional forms like monitoring temperature, blood pressure cuffs to glucose meters, pacemakers, defibrillators, and HIV testing.

New sensor applications are developing which produce individual, home, and environmental data. There are many sensor types that were developed years ago but are finding new applications. Navigational aids, such sensors as gyroscopes, accelerometers, and magnetometers, have existed for many years in flight instruments for aircraft and more modernly for smartphones. Sensors internal to smartphone devices are intended to monitor the device but can be repurposed to monitor to monitor many things such as extreme exposure to heat or movement for health applications. The interconnected network of devices to promote automation and efficiency is often referred to as the Internet of things (IoT). Sensors are becoming more prevalent and cheap enough that the public can make use of personal sensors that already exist in their daily lives or can be easily acquired.

Personal Health Monitoring

Health-monitoring applications are becoming increasingly common and produce very large volumes of data. Biophysical processes such as heart rate, breathing rate, sleep patterns, and restlessness can be recorded continuously using devices kept in contact with the body. Health-conscious and athletic communities, such as runners, have particularly taken to personal monitoring by using technology to track their current condition and progress. Pedometers, weight scales, and thermometers are commonplace. Heart rate, blood pressure, and muscle fatigue are now monitored

by affordable devices in the form of bracelets, rings, adhesive strips, and even clothing. Brands of smart clothing are offering built-in sensors for heart rate, respiration, skin temperature and moisture, and electrophysiological signals that are sometimes even recharged by solar panels. There are even wireless sensors for the insole of shoes to automatically adjust for the movements of the user in addition to providing health and training analysis.

Wearable health technologies are often used to provide individuals with private personal information; however, certain circumstances call for system-wide monitoring for medical or emergency purposes. Medical patients, such as those with diabetes or hypertension, can use continuously testing glucose meters or blood pressure monitors (Kalantar-zadeh 2013). Bluetooth-enabled devices can transmit data from monitoring sensors and contact the appropriate parties automatically if there are health concerns. Collective health information can be used to have a better understanding of such health concerns as cardiac issues, extreme temperatures, and even crisis information.

Smart Home

Sensors have long been a part of modern households from smoke and carbon monoxide detectors to security systems and motion sensors. Increasingly, smart home sensors are being used for everyday monitoring in order to have more efficient energy consumption with smart lighting fixtures and temperature controls. Sensors are often placed to inform on activities in the house such as a door or window being opened. This integrated network of house monitoring promises efficiency, automation, and safety based on personal preferences. There is significant investment in smart home technologies, and big data analysis can play a major role in determining appropriate settings based on feedback.

Environmental Monitoring

Monitoring of the environment from the surface to the atmosphere is traditionally a function performed by the government through remotely

sensed observations and broad surveys. Remote sensing imagery from satellites and airborne flights can create large datasets on global environmental changes for use in such applications as agriculture, pollution, water, climatic conditions, etc. Government agencies also employ static sensors and make on-site visits to check sensors which monitor environmental conditions. These sensors are sometimes integrated into networks which can communicate observations to form real-time monitoring systems.

In addition to traditional government sources of environmental data, there are growing collections of citizen science data that are focused primarily on areas of community concern such as air quality, water quality, and natural hazards. Air quality and water quality have long been monitored by communities concerned about pollution in their environment, but a recent development after the 2011 Fukushima nuclear disaster is radiation sensing. Safecast is a radiation monitoring project that seeks to empower people with information on environmental safety and openly distributes measurements under creative commons rights (McGrath and Scanail 2013). Radiation is not visibly observable so it is considered a “silent” environmental harm, and the risk needs to be considered in light of validated data (Hultquist and Cervone 2017). Citizen science projects for sensing natural hazards from flooding, landslides, earthquakes, wildfires, etc. have come online with support from both governments and communities. Open-source environmental data is a growing movement as people get engaged with their environment and become more educated about their health.

Conclusion

The development and availability of sensor technologies is a part of the big data paradigm. Sensors are able to produce an enormous amount of data, very quickly with real-time uploads, and from diverse types of sensors. Many questions still remain of how to use this data and if

connected sensors will lead to smart environments that will be a part of everyday modern life. The Internet of things (IoT) is envisioned to connect communication across domains and applications in order to enable the development of smart cities.

Sensor data can provide useful information for individuals and generalized information from collective monitoring. Services often offer personalized analysis in order to keep people engaged using the application. Yet, most analysis and interest from researchers in sensor data is at a generalized level. Despite mostly generalized data analysis, there is public concern related to data privacy from individual and home sensors. The privacy level of the data is highly dependent on the system used and the terms of service agreement if a service is being provided related to the sensor data.

Analysis of sensor data is often complex, messy, and hard to verify. Nonpersonal data can often be checked or referenced to a comparable dataset to see if it makes sense. However, large datasets produced by personal sensors for such applications as health are difficult to independently verify at an individual level. For example, an environmental condition could have caused a natural reaction of a rapid heartbeat which is medically safe given the condition that the user awoke with a quick increase in heart rate due to an earthquake. Individual inspection of data for such noise is fraught with problems as it is complicated to identify causes in the raw data from an individual, but at a generalized level, such data can be valuable for research and can appropriately take into account variations in the data.

Sensor technologies are integrated into everyday life and are used in numerous applications to monitor conditions. The usefulness of technological sensors should be no surprise as every living organism has biological sensors which serve similar purposes to indicate the regulation of internal functions and conditions of the external environment. The integration of sensor technologies is a natural step that goes from individual measurements to collective monitoring which highlights the need for big data analysis and validation.

Cross-References

- [AgInformatics](#)
- [Biometrics](#)
- [Biosurveillance](#)
- [Crowdsourcing](#)
- [Drones](#)
- [Environment](#)
- [Health Informatics](#)
- [Participatory Health and Big Data](#)
- [Patient-Centered \(Personalized\) Health](#)
- [Pollution, Air](#)
- [Pollution, Land](#)
- [Pollution, Water](#)
- [Satellite Imagery/Remote Sensing](#)

Further Reading

- Hultquist, C., & Cervone, G. (2017). Citizen monitoring during hazards: Validation of Fukushima radiation measurements. *Geo Journal*. <http://doi.org/10.1007/s10708-017-9767-x>.
- Kalantar-zadeh, K. (2013). *Sensors: An introductory course* (1st ed.). Boston: Springer US.
- McGrath, M. J., & Scanail, C. N. (2013). *Sensor technologies: Healthcare, wellness, and environmental applications*. New York: Apress Open.

Sentic Computing

Erik Cambria

School of Computer Science and Engineering,
Nanyang Technological University, Singapore,
Singapore

With the recent development of deep learning, research in artificial intelligence (AI) has gained new vigor and prominence. Machine learning, however, suffers from three big issues, namely:

1. Dependency issue: it requires (a lot of) training data and it is domain-dependent.
2. Consistency issue: different training and/or tweaking lead to different results.

3. Transparency issue: the reasoning process is uninterpretable (black-box algorithms).

Sentic computing (Cambria and Hussain 2015) addresses these issues in the context of natural language processing (NLP) by coupling machine learning with linguistics and commonsense reasoning. In particular, we apply an ensemble of commonsense-driven linguistic patterns and statistical NLP: the former are triggered when prior knowledge is available, the latter is used as backup plan when both semantics and sentence structure are unknown. Machine learning, in fact, is only useful to make a *good guess* because it only encodes correlation and its decision-making process is merely probabilistic. To use Noam Chomsky’s words, “you do not get discoveries in the sciences by taking huge amounts of data, throwing them into a computer and doing statistical analysis of them: that’s not the way you understand things, you have to have theoretical insights.”

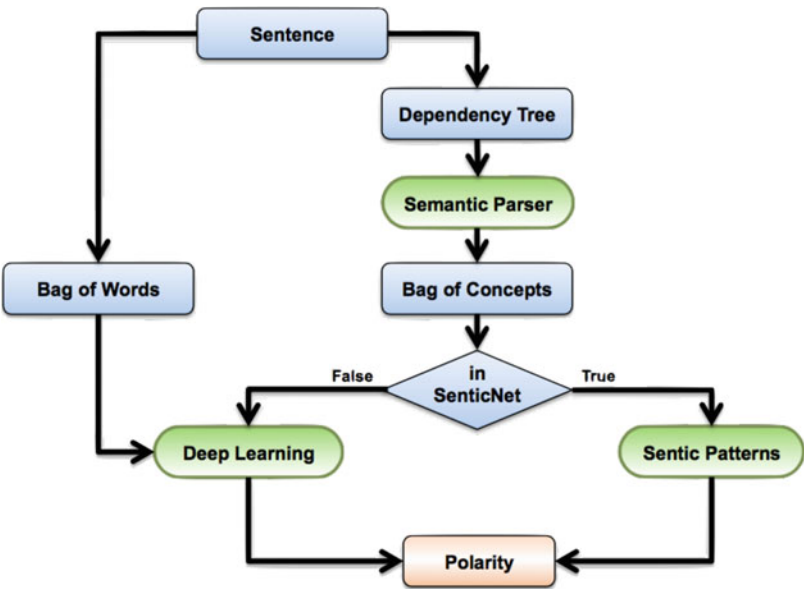
Sentic computing is a multidisciplinary approach to natural language understanding that aims to bridge the gap between statistical NLP and many other disciplines that are necessary for understanding human language, such as linguistics, commonsense reasoning, affective

computing, and more. Sentic computing, whose term derives from the Latin “sensus” (as in commonsense) and “sentire” (root of words such as sentiment and sentence), enables the analysis of text not only at document, page, or paragraph level, but also at sentence, clause, and concept level (Fig. 1).

Sentic computing positions itself as a horizontal technology that serves as a back-end to many different applications in the areas of e-business, e-commerce, e-governance, e-security, e-health, e-learning, e-tourism, e-mobility, e-entertainment, and more. Some examples of such applications include financial forecasting (Xing et al. 2018) and healthcare quality assessment (Cambria et al. 2012a), community detection (Cavallari et al. 2017) and cyber issue detection (Cambria et al. 2010), human communication comprehension (Zadeh et al. 2018) and dialogue systems (Young et al. 2018). State-of-the-art performance is ensured in all these sentiment analysis applications, thanks to sentic computing’s new approach to NLP, whose novelty gravitates around three key shifts:

- 1. Shift from mono- to multidisciplinary – evidenced by the concomitant use of AI and Semantic Web techniques, for knowledge representation and inference; mathematics, for

Sentic Computing,
Fig. 1 Sentic computing flowchart



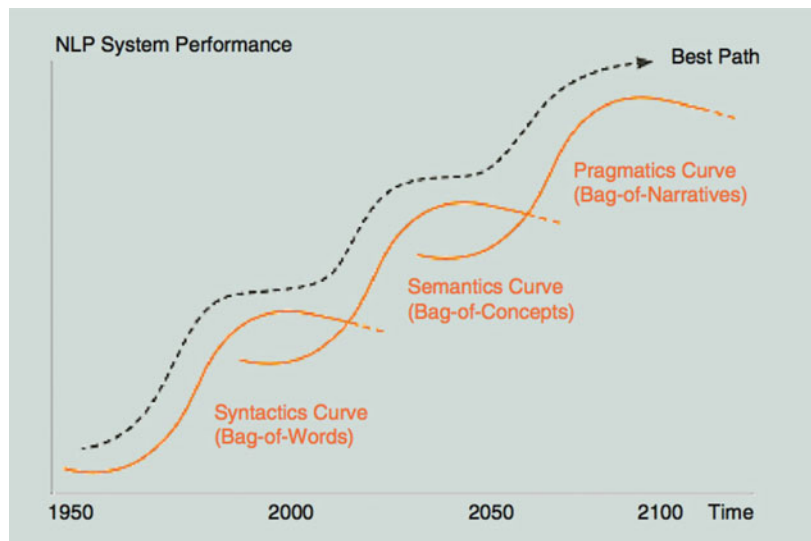
carrying out tasks such as graph mining and multidimensionality reduction; linguistics, for discourse analysis and pragmatics; psychology, for cognitive and affective modeling; sociology, for understanding social network dynamics and social influence; finally ethics, for understanding related issues about the nature of mind and the creation of emotional machines.

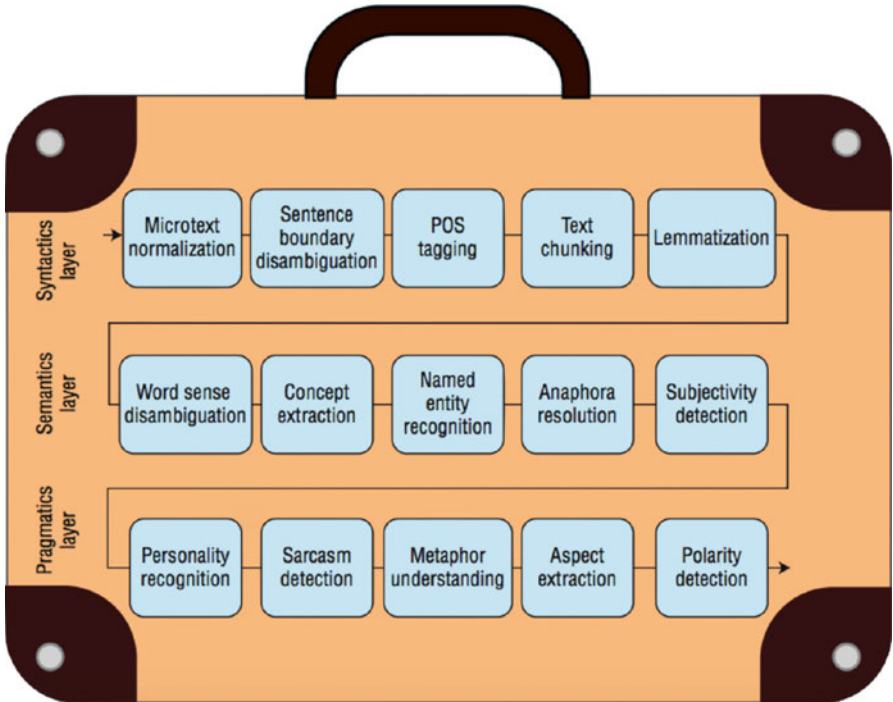
2. Shift from syntax to semantics – enabled by the adoption of the bag-of-concepts model instead of simply counting word co-occurrence frequencies in text. Working at concept-level entails preserving the meaning carried by multiword expressions such as cloud computing, which represent “semantic atoms” that should never be broken down into single words. In the bag-of-words model, for example, the concept cloud computing would be split into computing and cloud, which may wrongly activate concepts related to the weather and, hence, compromise categorization accuracy.
3. Shift from statistics to linguistics – implemented by allowing sentiments to flow from concept to concept based on the dependency relation between clauses. The sentence “iPhoneX is expensive but nice”, for example, is equal to “iPhoneX is nice but expensive” from a bag-of-words perspective. However,

the two sentences bear opposite polarity: the former is positive as the user seems to be willing to make the effort to buy the product despite its high price; the latter is negative as the user complains about the price of iPhoneX although he/she likes it (Fig. 2).

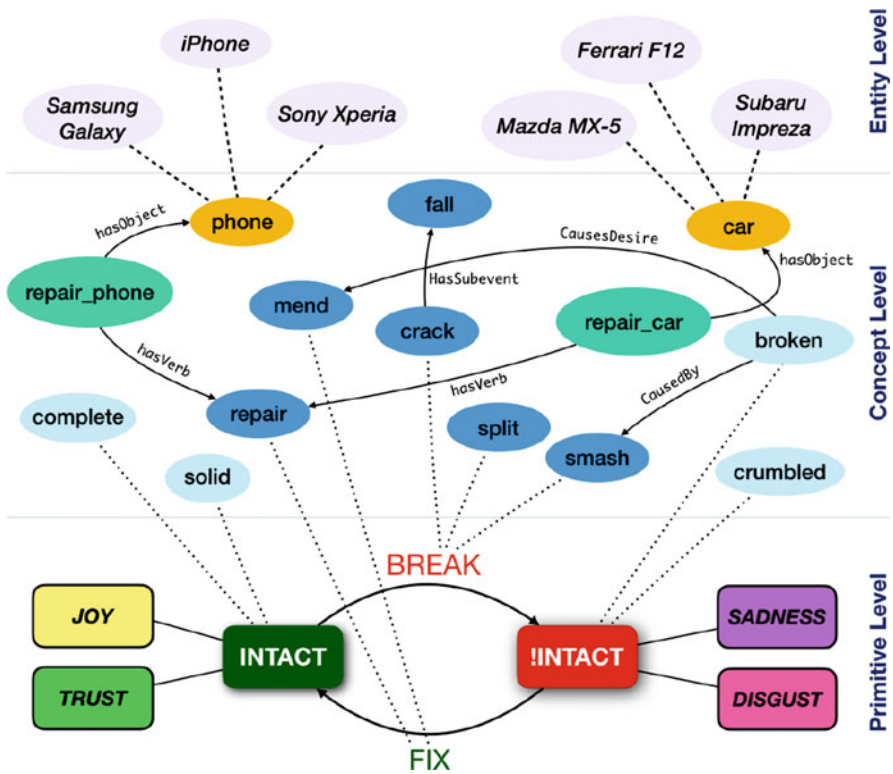
Sentic computing takes a holistic approach to natural language understanding by handling the many subproblems involved in extracting meaning and polarity from text. While most works approach it as a simple categorization problem, in fact, sentiment analysis is actually a suitcase research problem (Cambria et al. 2017b) that requires tackling many NLP tasks (Fig. 3). As Marvin Minsky would say, the expression “sentiment analysis” itself is a big suitcase (like many others related to affective computing (Cambria et al. 2017a), e.g., emotion recognition or opinion mining) that all of us use to encapsulate our jumbled idea about how our minds convey emotions and opinions through natural language. Sentic computing addresses the composite nature of the problem via a three-layer structure that concomitantly handles tasks such as subjectivity detection (Chaturvedi et al. 2018), to filter out neutral content, named-entity recognition (Ma et al. 2016), to locate and classify named entities into pre-defined categories, personality recognition (Majumder et al. 2017), for distinguishing between different

Sentic Computing,
Fig. 2 Jumping NLP
curves

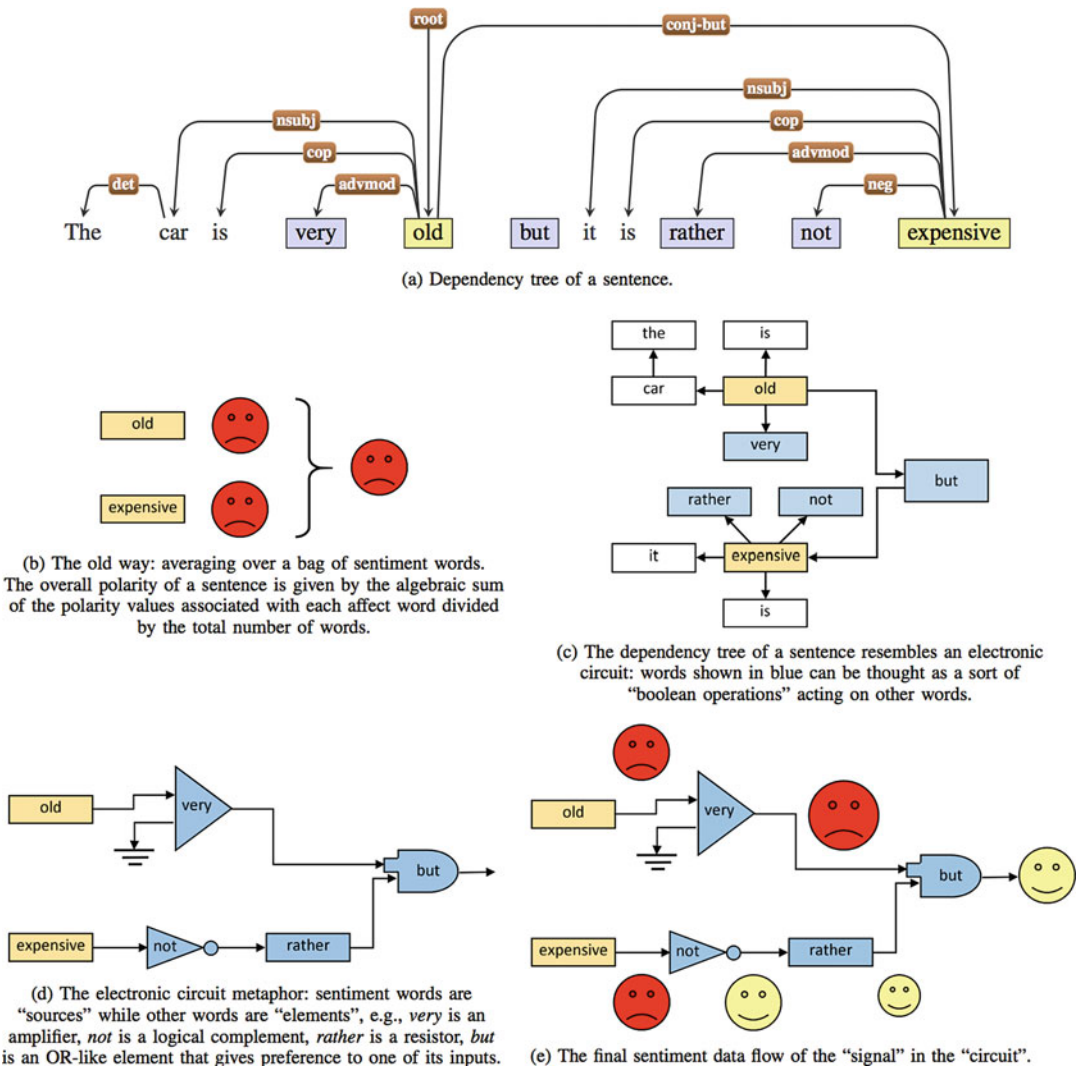




Sentic Computing, Fig. 3 Sentiment analysis suitcase



Sentic Computing, Fig. 4 SenticNet



Sentic Computing, Fig. 5 Sentic patterns

personality types of the users, sarcasm detection (Poria et al. 2016), to detect and handle sarcasm in opinions, aspect extraction (Ma et al. 2018), for enabling aspect-based sentiment analysis, and more.

The core element of sentic computing is SenticNet (Cambria et al. 2020), a knowledge base of 200,000 commonsense concepts (Fig. 4). Unlike many other sentiment analysis resources, SenticNet is not built by manually labeling pieces of knowledge coming from general NLP resources such as WordNet or DBpedia. Instead, it is automatically constructed by applying graph-

mining and multidimensional scaling techniques on the affective commonsense knowledge collected from three different sources, namely: WordNet-Affect, Open Mind Common Sense, and a game engine for commonsense knowledge acquisition (GECKA) (Cambria et al. 2015b). This knowledge is represented redundantly at three levels (following Minsky’s panalogy principle): semantic network, matrix, and vector space (Cambria et al. 2015a). Subsequently, semantics and sentics are calculated though the ensemble application of spreading activation (Cambria et al. 2012c), neural networks (Ma et al. 2018),

and an emotion categorization model (Susanto et al. 2020).

While SenticNet can be used as any other sentiment lexicon, e.g., concept matching or bag-of-concepts model, the right way to use the knowledge base for the task of polarity detection is in conjunction with sentic patterns (Poria et al. 2014). Sentic patterns are sentiment-specific linguistic patterns that infer polarity by allowing affective information to flow from concept to concept based on the dependency relation between clauses. The main idea behind such patterns can be best illustrated by analogy with an electronic circuit, in which few “elements” are “sources” of the charge or signal, while many elements operate on the signal by transforming it or combining different signals. This implements a rudimentary type of semantic processing, where the “meaning” of a sentence is reduced to only one value: its polarity.

Sentic patterns are applied to the dependency syntactic tree of a sentence, as shown in Fig. 5a. The only two words that have intrinsic polarity are shown in yellow color; the words that modify the meaning of other words in the manner similar to contextual valence shifters are shown in blue. A baseline that completely ignores sentence structure, as well as words that have no intrinsic polarity, is shown in Fig. 5b: the only two words left are negative and, hence, the total polarity is negative. However, the syntactic tree can be reinterpreted in the form of a “circuit” where the “signal” flows from one element (or subtree) to another, as shown in Fig. 5c. After removing the words not used for polarity calculation (in white), a circuit with elements resembling electronic amplifiers, logical complements, and resistors is obtained, as shown in Fig. 5d.

Figure 5e illustrates the idea at work: the sentiment flows from polarity words through shifters and combining words. The two polarity-bearing words in this example are negative. The negative effect of the word “old” is amplified by the intensifier “very”. However, the negative effect of the word “expensive” is inverted by the negation, and the resulting positive value is decreased by the “resistor”. Finally, the values of the two phrases are combined by the conjunction “but”, so that the

overall polarity has the same sign as that of the second component (positive).

Further Reading

- Cambria, E., & Hussain, A. (2015). *Sentic computing: A common-sense-based framework for concept-level sentiment analysis*. Cham: Springer.
- Cambria, E., Chandra, P., Sharma, A., & Hussain, A. (2010). Do not feel the trolls. In *ISWC*. Shanghai
- Cambria, E., Benson, T., Eckl, C., & Hussain, A. (2012a). Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Systems with Applications*, 39(12), 10533–10543.
- Cambria, E., Livingstone, A., & Hussain, A. (2012b). The hourglass of emotions. In A. Esposito, A. Vinciarelli, R. Hoffmann, & V. Muller (Eds.), *Cognitive behavioral systems, Lecture notes in computer science* (Vol. 7403, pp. 144–157). Berlin/Heidelberg: Springer.
- Cambria, E., Olsher, D., & Kwok, K. (2012c). Sentic activation: A two-level affective common sense reasoning framework. In *AAAI* (pp. 186–192). Toronto.
- Cambria, E., Fu, J., Bisio, F., & Poria, S. (2015a). AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. In *AAAI* (pp. 508–514). Austin.
- Cambria, E., Rajagopal, D., Kwok, K., & Sepulveda, J. (2015b). GECKA: Game engine for commonsense knowledge acquisition. In *FLAIRS* (pp. 282–287).
- Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017a). *A practical guide to sentiment analysis*. Cham: Springer.
- Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017b). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6), 74–80.
- Cambria, E., Li, Y., Xing, Z., Poria, S., & Kwok, K. (2020). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *CIKM*. Ireland.
- Cavallari, S., Zheng, V., Cai, H., Chang, K., & Cambria, E. (2017). Learning community embedding with community detection and node embedding on graphs. In *CIKM* (pp. 377–386). Singapore.
- Chaturvedi, I., Ragusa, E., Gastaldo, P., Zunino, R., & Cambria, E. (2018). Bayesian network based extreme learning machine for subjectivity detection. *Journal of The Franklin Institute*, 355(4), 1780–1797.
- Ma, Y., Cambria, E., & Gao, S. (2016). Label embedding for zero-shot fine-grained named entity typing. In *COLING* (pp. 171–180). Osaka.
- Ma, Y., Peng, H., & Cambria, E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *AAAI* (pp. 5876–5883). New Orleans.
- Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for

- personality detection from text. *IEEE Intelligent Systems*, 32(2), 74–79.
- Poria, S., Cambria, E., Winterstein, G., & Huang, G.-B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69, 45–63.
- Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. In *COLING* (pp. 1601–1612). Osaka.
- Susanto, Y., Livingstone, A., Ng, B.C., & Cambria, E. (2020) The Hourglass model revisited. *IEEE Intelligent Systems* 35(5).
- Xing, F., Cambria, E., & Welsch, R. (2018). Natural language based financial forecasting: A survey. *Artificial Intelligence Review*. <https://doi.org/10.1007/s10462-017-9588-9>.
- Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., & Huang, M. (2018). Augmenting end-to-end dialog systems with commonsense knowledge. In *AAAI* (pp. 4970–4977). New Orleans.
- Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L.-P. (2018). Multi-attention recurrent network for human communication comprehension. In *AAAI* (pp. 5642–5649). New Orleans.

Sentiment Analysis

Francis Dalisay¹, Matthew J. Kushin² and Masahiro Yamamoto³

¹Communication & Fine Arts, College of Liberal Arts & Social Sciences, University of Guam, Mangilao, GU, USA

²Department of Communication, Shepherd University, Shepherdstown, WV, USA

³Department of Communication, University at Albany – SUNY, Albany, NY, USA

Sentiment analysis is defined as the computational study of opinions, or sentiment, in text. Sentiment analysis typically intends to capture an opinion holder's evaluative response (e.g., positive, negative, or neutral, or a more fine-grained classification scheme) toward an object. The evaluative response reflects an opinion holder's attitudes, or affective feelings, beliefs, thoughts, and appraisals.

Francis Dalisay, Matthew Kushin, and Masahiro Yamamoto contributed equally to the writing of this entry.

According to scholars Erik Cambria, Bjorn Schuller, Yunging Xia, and Catherine Havasi, sentiment analysis is a term typically used interchangeably with opinion mining to refer to the same field of study. The scholars note, however, that opinion mining generally involves the detection of the polarity of opinion, also referred to as the sentiment orientation of a given text (i.e., whether the expressed opinion is positive, negative, or neutral). Sentiment analysis focuses on the recognition of emotion (e.g., emotional states such as “sad” or “happy”), but also typically involves some form of opinion mining. For this reason, and since both fields rely on natural language processing (NLP) to analyze opinions from text, sentiment analysis is often couched under the same umbrella as opinion mining.

Sentiment analysis has gained popularity as a social data analytics tool. Recent years have witnessed the widespread adoption of social media platforms as outlets to publicly express opinions on nearly any subject, including those relating to political and social issues, sporting and entertainment events, weather, and brand and consumer experiences. Much of the content posted on sites such as Twitter, Facebook, YouTube, customer review pages, and news article comment boards is public. As such, businesses, political campaigns, universities, and government entities, among others, can collect and analyze this information to gain insight into the thoughts of key publics.

The ability of sentiment analysis to measure individuals' thoughts and feelings has a wide range of practical applications. For example, sentiment analysis can be used to analyze online news content and to examine the polarity of news coverage of particular issues. Also, businesses are able to collect and analyze the sentiment of comments posted online to assess consumers' opinions toward their products and services, evaluate the effectiveness of advertising and PR campaigns, and identify customer complaints. Gathering such market intelligence helps guide decision-making in the realms of product research and development, marketing and public relations, crisis management, and

customer relations. Although businesses have traditionally relied on surveys and focus groups, sentiment analysis offers several unique advantages over such conventional data collection methods. These advantages include reduced cost and time, increased access to much larger samples and hard-to-reach populations, and real-time intelligence. Thus, sentiment analysis can be a useful market research tool. Indeed, sentiment analysis is now commonly offered by many commercial social data analysis services.

Approaches

Broadly speaking, there exist two approaches in the automatic extraction of sentiment from textual material: the lexicon-based approach and the machine learning-based approach. In the lexicon-based approach, a sentiment orientation score is calculated for a given text unit based on a predetermined set of opinion words with positive (e.g., good, fun, exciting) and negative (e.g., bad, boring, poor) sentiments. In a simple form, a list of words, phrases, and idioms with known sentiment orientations is built into a dictionary, or an opinion lexicon. Each word is assigned specific sentiment orientation scores. Using the lexicon, each opinion word extracted receives a predefined sentiment orientation score, which is then aggregated for a text unit.

The machine learning-based approach, also called the text classification approach, builds a sentiment classifier to determine whether a given text about an object is positive, negative, or neutral. Using the ability of machines to learn, this approach trains a sentiment classifier to use a large set of examples, or training corpus, that have sentiment categories (e.g., positive, negative, or neutral). The sentiment categories are manually annotated by humans according to predefined rules. The classifier then applies the properties of the training corpus to classify data into sentiment categories.

Levels of Analysis

The classification of an opinion in text as positive, negative, or neutral (or a more fine-grained classification scheme) is impacted by and thus requires consideration of the level at which the analysis is conducted. There are three levels of analysis: document, sentence, and aspect and/or entity. First, the document-level sentiment classification addresses a whole document as the unit of analysis. The task of this level of analysis is to determine whether an entire document (e.g., a product review, a blog post, an email, etc.) is positive, negative, or neutral about an object. This level of analysis assumes that the opinions expressed on the document are targeted toward a single entity (e.g., a single product). As such, this level is not particularly useful to documents that discuss multiple entities.

The second, sentence-level sentiment classification, focuses on the sentiment orientation of individual sentences. This level of analysis is also referred to as subjectivity classification and comprised of two tasks: subjective classification and sentence-level classification. In the first task, the system determines whether a sentence is subjective or objective. If it is determined that the sentence expresses a subjective opinion, the analysis moves to the second task, sentence-level classification. This second task involves determining whether the sentence is positive, negative, or neutral.

The third type of classification is referred to as entity and aspect-level sentiment analysis. Also called feature-based opinion mining, this level of analysis focuses on sentiments directed at entities and/or their aspects. An entity can include a product, service, person, issue, or event. An aspect is a feature of the entity, such as its color or weight. For example, in the sentence “the design of this laptop is bad, but its processing speed is excellent,” there are two aspects stated – “design” and “processing speed.” This sentence is negative about one aspect, “design,” and positive about the other aspect, “processing speed.” Entity- and aspect-level sentiment analysis is not limited to

analyzing documents or sentences alone. Indeed, although a document or sentence may contain opinions regarding multiple entities and their aspects, the entity- and aspect-level sentiment analysis has the ability to identify the specific entities and/or aspects that the opinions on the document or sentence are referring to and then determine whether the opinions are positive, negative, or neutral.

Challenges and Limitations

Extracting opinions from texts is a daunting task. It requires a thorough understanding of the semantic, syntactic, explicit, and implicit rules of a language. Also, because sentiment analysis is carried out by a computer system with a typical focus on analyzing documents on a particular topic, off-topic passages containing irrelevant information may also be included in the analyses (e.g., a document may contain information on multiple topics). This could result in creating inaccurate global sentiment polarities about the main topic being analyzed. Therefore, the computer system must be able to adequately screen and distinguish opinions that are not relevant to the topic being analyzed. Relatedly, for the machine learning-based approach, a sentiment classifier trained on a certain domain (e.g., car reviews) may perform well on the particular topic, but may not when applied to another domain (e.g., computer review). The issue of domain independence is another important challenge.

Also, the complexities of human communication limit the capacity of sentiment analysis to capture nuanced, contextual meanings that opinion holders actually intend to communicate in their messages. Examples include the use of sarcasm, irony, and humor in which context plays a key role in conveying the intended message, particularly in cases when an individual says one thing but means the opposite. For example, someone may say “nice shirt,” which implies positive sentiment if said sincerely but implies negative

sentiment if said sarcastically. Similarly, words such as “sick,” “bad,” and “nasty” may have reversed sentiment orientation depending on context and how they are used. For example, “My new car is sick!” implies positive sentiment toward the car. These issues can also contribute to inaccuracies in sentiment analysis.

Altogether, despite these limitations, the computational study of opinions provided by sentiment analysis can be beneficial for practical purposes. So long as individuals continue to share their opinions through online user-generated media, the possibilities for entities seeking to gain meaningful insights into the opinions of key publics will remain. Yet, challenges to sentiment analysis such as those discussed above, pose significant limitations to its accuracy and thus its usefulness in decision-making.

Cross-References

- [Brand Monitoring](#)
- [Data Mining](#)
- [Facebook](#)
- [LinkedIn](#)
- [Online Advertising](#)
- [Online Identity](#)
- [SalesForce](#)
- [Social Media](#)
- [Time Series Analytics](#)

Further Reading

- Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28, 15–21.
- Liu, B. (2011). *Sentiment analysis and opinion mining*. San Rafael: Morgan & Claypool.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.
- Pang, B., Lee, L., & Vaithyanathan S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 79–86).

Zezima, K. The secret service wants software that detects sarcasm (Yeah, good luck.) *The Washington Post*. Retrieved 11 Aug 2014 from http://www.washingtonpost.com/politics/the-secret-service-wants-software-that-detects-sarcasm-yeah-good-luck/2014/06/03/35bb8bd0-eb41-11e3-9f5c-9075d5508f0a_story.html.

Server Farm

► [Data Center](#)

Silviculture

► [Forestry](#)

“Small” Data

Rochelle E. Tractenberg^{1,2} and
Kimberly F. Sellers³

¹Collaborative for Research on Outcomes and Metrics, Washington, DC, USA

²Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and Rehabilitation Medicine, Georgetown University, Washington, DC, USA

³Department of Mathematics and Statistics, Georgetown University, Washington, DC, USA

Synonyms

[Data](#); [Statistics](#)

Introduction

Big data are often characterized by “the 3 Vs”: volume, velocity, and variety. This implies that “small data” lack these qualities, but that is an incorrect conclusion about what defines “small” data. Instead, we define “small data” to be simply “data” – specifically, data that are finite but not

necessarily “small” in scope, dimension, or rate of accumulation. The characterization of data as “small” is essentially dependent on the context and use for which the data are intended. In fact, disciplinary perspectives vary on how large “big data” need to be to merit this label, but small data are not characterized effectively by the *absence* of one or more of these “3 Vs.” Most statistical analyses require some amount of vector and matrix manipulation for efficient computation in the modern context. Data sets may be considered “big” if they are so large, multidimensional, and/or quickly accumulating in size that the typical linear algebraic manipulations cannot converge or yield true summaries of the full data set. The fundamental statistical analyses, however, are the same for data that are “big” or “small”; the true distinction arises from the extent to which computational manipulation is required to map and reduce the data (Day and Ghemawat 2004) such that a coherent result can be derived. All analyses share common features, irrespective of the size, complexity, or completeness of the data – the relationship between statistics and the underlying population; the association between inference, estimation, and prediction; and the dependence of interpretation and decision-making on statistical inference. To expand on the lack of distinguishability between “small” data and “big” data, we explore each of these features in turn. By doing so, we expound on the assertion that a characterization of a dataset as “small” depends on the users’ intention and the context in which the data, and results from its analysis, will be used.

Understanding “Big Data” as “Data”

An understanding of why some datasets are characterized as “big” and/or “small” requires some juxtaposition of these two descriptors. “Big data” are thought to expand the boundary of data science because innovation has been ongoing to promote ever-increasing capacity to collect and analyze data with high volume, velocity, and/or variety (i.e., the 3 Vs). In this era of technological advances, computers are able to maintain and

process terabytes of information, including records, transactions, tables, files, etc. However, the ability to analyze data has *always* depended on the methodologies, tools, and technology available at the time; thus the reliance on computational power to collect or process data is not new or specific to the current era and cannot be considered to delimit "big" from "small" data.

Data collection and analyses date back to ancient Egyptian civilizations that collected census information; the earliest Confucian societies collected this population-spanning data as well. These efforts were conducted by hand for centuries, until a "tabulating machine" was used to complete the analyses required for the 1890 United States Census; this is possibly the first time so large a dataset was analyzed with a non-human "computer." Investigations that previously took years to achieve were suddenly completed in a fraction of the time (months!). Since then, technology continues to be harnessed to facilitate data collection, management, and analysis. In fact, when it was suggested to add "data science" to the field of statistics (Bickel 2000; Rao 2001), "big data" may have referred to a data set of up to several gigabytes in size; today, petabytes of data are not uncommon. Therefore, neither the size nor the need for technological advancements are inherent properties of either "big" or "small" data.

Data are sometimes called "big" if the data collection process is fast(-er), not finite in time or amount, and/or inclusive of a wide range of formats and quality. These features may be contrasted with experimental, survey, epidemiologic, or census data where the data structure, timing, and format are fixed and typically finite. Technological advances allow investigators to collect batches of experimental, survey, or other traditional types of data in near-real or real time, or in online or streaming fashion; such information has been incorporated to ask and answer experimental and epidemiologic questions, including testing hypotheses in physics, climate, chemistry, and both social and biomedical sciences, since the technology was developed. It is inappropriate to distinguish "big" from "small"

data along these characteristics; in fact, two analysts simultaneously considering the same data set may each perceive it to be "big" or "small"; these labels must be considered to be *relative*.

Analysis and Interpretation of "Big Data" Is Based on Methods for "Small Data"

Considering analysis, manipulation, and interpretation of data can support a deeper appreciation for the differences and similarities of "big" and "small" data. Large(r) and higher-dimensional data sets may require computational manipulation (e.g., Day and Ghemawat 2004), including grouping and dimension reduction, to derive an interpretable result from the full data set. Further, whenever a larger/higher dimension dataset is partitioned for analysis, the partitions or subsets are analyzed using standard statistical methods. The following sections explicate how standard statistical analytic methods (i.e., for "small" data) are applied to a dataset whether it is described as "small" or "big". These methods are selected, employed, and interpreted specifically to support the user's intention for the results and do not depend inherently on the size or complexity of the data itself. This underscores the difficulty of articulating any specific criterion/a for characterizing data as "big" or "small."

Sample Versus Population

Statistical analysis and summarization of "big" data are the same as for data generally; the description, confidence/uncertainty, and coherence of the results may vary with the size and completeness of the data set. Even the largest and most multidimensional dataset is presumably an incomplete (albeit massive) representation of the entire universe of values – the "population." Thus, the field of statistics has historically been based on long-run frequencies or computed estimates of the true population parameters. For example, in some current massive data collection and warehousing enterprises, the full population can never be obtained because the data are

continuously streaming in and collected. In other massive data sets, however, the entire population *is* captured; examples include the medical records for a health insurance company, sales on [Amazon.com](https://www.amazon.com), or weather data for the detection of an evolving storm or other significant weather pattern. The fundamental statistical analyses would be the same for either of these data types; however, they would result in *estimates* for the (essentially) infinite data set, while actual population-descriptive values are possible whenever finite/population data are obtained. Importantly, it is not the size or complexity of the data that results in either estimation or population description – it is whether or not the data are finite. This underscores the reliance of any and all data analysis procedures on statistical methodologies; assumptions about the data are required for the correct use and interpretation of these methodologies for data of any size and complexity. It further blurs qualifications of a given data set as “big” or “small.”

Inference, Estimation, and Prediction

Statistical methods are generally used for two purposes: (1) to estimate “true” population parameters when only sample information is available, and (2) to make or test predictions about either future results or about relationships among variables. These methods are used to infer “the truth” from incomplete data and are the foundations of nearly all experimental designs and tests of quantitative hypotheses in applied disciplines (e.g., science, engineering, and business). Modern statistical analysis generates results (i.e., parameter estimates and tests of inferences) that can be characterized with respect to how rare they are given the random variability inherent in the data set.

In frequentist statistical analysis (based on long run results), this characterization typically describes how likely the observed result would be if there were, in truth, no relationship between (any) variables, or if the true parameter value was a specific value (e.g., zero). In Bayesian statistical analysis (based on current data and prior knowledge), this characterization describes how likely it

is that there is truly no relationship given the data that were observed and prior knowledge about whether such a relationship exists.

Whenever inferences are made about estimates and predictions about future events, relationships, or other unknown/unobserved events or results, corrections must be made for the multitude of inferences that are made for both frequentist and Bayesian methods. Confidence and uncertainty about every inference and estimate must accommodate the fact that more than one has been made; these “multiple comparisons corrections” protect against decisions that some outcome or result is rare/statistically significant when, in fact, the variability inherent in the data make that result far less rare than it appears. Numerous correction methods exist with modern (since the mid-1990s) approaches focusing not on controlling for “multiple comparisons” (which are closely tied to experimental design and formal hypothesis testing), but controlling the “false discovery rate” (which is the rate at which relationships or estimates will be declared “rare given the inherent variability of the data” when they are not, in fact, rare). Decisions made about inferences, estimates, and predictions are classified as correct (i.e., the event is rare and is declared rare, or the event is not rare and is declared not rare) or incorrect (i.e., the event is rare but is declared not rare – a false negative/Type II error; or the event is not rare but is declared rare – a false positive/Type I error); controls for multiple comparisons or false discoveries seek to limit Type I errors.

Decisions are made based on the data analysis, which holds for “big” or “small” data. While multiple comparisons corrections and false discovery rate controls have long been accepted as representing competent scientific practice, they are also essential features of the analysis of big data, whether or not these data are analyzed for scientific or research purposes.

Analysis, Interpretation, and Decision Making

Analyses of data are either motivated by theory or prior evidence (“theory-driven”), or they are unplanned and motivated by the data themselves

(“data-driven”). Both types of investigations can be executed on data of any size, complexity, or completeness. While the motivations for data analysis vary across disciplines, evidence that supports decisions is always important. Statistical methods have been developed, validated, and utilized to support the most appropriate analysis, given the data and its properties, so that defensible and reproducible interpretations and inferences result. Thus, decisions that are made based on the analysis of data, whether “big” or “small,” are inherently dependent on the quality of the analysis and associated interpretations.

Conclusion

As has been the case for centuries, today’s “big” data will eventually be perceived as “small”; however, the statistical methodologies for analyzing and interpreting all data will also continue to evolve, and these will become increasingly interdependent on the methods for collecting, manipulating, and storing the data. Because of the constant evolution and advancement in technology and computation, the notion of “big data” may be best conceptualized as representing the *processes* of data collection, storage, and manipulation for interpretable analysis, and not the size, utility, or complexity of the data itself. Therefore, the characterization of data as “small” depends critically on the context and use for which the data are intended.

Further Reading

- Bickel, P. J. (2000). Statistics as the information science. *Opportunities for the mathematical sciences*, 9, 11.
- Day, J., & Ghemawat, S (2004, December). MapReduce: Simplified data processing on large clusters. In *OSDI'04: Sixth symposium on operating system design and implementation*. San Francisco. Downloaded from <https://research.google.com/archive/mapreduce.html> on 21 Dec 2016.
- Rao, C. R. (2001). Statistics: Reflections on the past and visions for the future. *Communications in Statistics – Theory and Methods*, 30(11), 2235–2257.

Smart Agriculture

► Agriculture

Smart Cities

Jan Lauren Boyles
Greenlee School of Journalism and
Communication, Iowa State University, Ames,
IA, USA

Definition/Introduction

Smart cities are built upon aggregated, data-driven insights that are obtained directly from the urban infrastructure. These data points translate into actionable information that can guide municipal development and policy (Albino et al. 2015). Building on the emergent Internet of Things movement, networked sensors (often physically embedded into the built environment) create rich data streams that uncover how city resources are used (Townsend 2013; Komninos 2015; Sadowski and Pasquale 2015). Such intelligent systems, for instance, can send alerts to city residents when demand for urban resources outpaces supply or when emergency conditions exist within city limits. By analyzing these data flows (often in real time), elected officials, city staff, civic leaders, and average citizens can more fully understand resource use and allocation, thereby optimizing the full potential of municipal services (Hollands 2008; de Lange and de Waal 2013; Campbell 2013; Komninos 2015). Over time, the integration of such intelligent systems into metropolitan life acts to better inform urban policy making and better direct long-term municipal planning efforts (Batty 2013; Komninos 2015; Goldsmith and Crawford 2014). Despite this promise of more effective and responsive governance, however, achieving a truly smart city often requires the redesign (and in many

cases, the physical rebuilding) of structures to harvest and process big data from the urban environment (Campbell 2013). As a result, global metropolitan leaders continue to experiment with cost-effective approaches to constructing smart cities in the late-2010s.

Heralded as potentially revolutionizing citizen-government interactions within cities, the initial integration of Internet Communication Technologies (ICTs) into the physical city in the late 1990s was viewed as the first step toward today's smart cities (Caragliu et al. 2011; Albino et al. 2015). In the early 2000s, the burgeoning population growth of global cities mandated the use of more sophisticated computational tools to effectively monitor and manage metropolitan resources (Campbell 2013; Meijer and Bolivar 2015). The rise of smart cities in the early 2010s can, in fact, be traced to a trio of technological advances: the adoption of cloud computing, the expansion of wireless networks, and the acceleration of processing power. At the same time, the societal uptick in mobile computing by everyday citizens enables more data to be collected on user habits and behaviors of urban residents (Batty 2013). The most significant advance in smart city adoption rests, however, in geolocation – the concept that data can be linked to physical space (Batty 2013; Townsend 2013). European metropolises, in particular, have been early adopters of intelligent systems (Vanolo 2013).

The Challenges of Intelligent Governance

Tactically, most smart cities attempt to tackle *wicked problems* – the types of dilemmas that have historically puzzled city planners (Campbell 2013; Komninos 2015). The integration of intelligent systems into the urban environment has accelerated the time horizon for policymaking for these issues (Batty 2013). Data that once took years to gather and assess can now be accumulated and analyzed in mere hours, or in some cases, in real time (Batty 2013). Within smart cities, crowdsourcing efforts often also enlist residents, who voluntarily provide data to fuel collective

and collaborative solutions (Batty 2013). Operating in this environment of heightened responsiveness, municipal leaders within smart cities are increasingly expected to integrate open data initiatives that provide public access to the information gathered by the data-driven municipal networks (Schrock 2016). City planners, civic activists, and urban technologists must also jointly consider the needs of city dwellers throughout the process of designing smart cities, directly engaging residents in the building of smart systems (de Lange and de Waal 2013). At the same time, urban officials must be increasingly cognizant that as more user behaviors within city limits are tracked with data, the surveillance required to power smart systems may also concurrently challenge citizen notions of privacy and security (Goldsmith and Crawford 2014; Sadowski and Pasquale 2015). Local governments must also ensure that the data collected will be safe and secure from hackers, who may wish to disrupt essential smart systems within cities (Schrock 2016).

Conclusion

The successful integration of intelligent systems into the city is centrally predicated upon financial investment in overhauling aging urban infrastructure (Townsend 2013; Sadowski and Pasquale 2015). Politically, investment decisions are further complicated by fragmented municipal leadership, whose priorities for smart city implementation may shift between election cycles and administrations (Campbell 2013). Rather than encountering these challenges in isolation, municipal leaders are beginning to work together to develop global solutions to shared wicked problems. Intelligent system advocates argue that developing collaborative approaches to building smart cities will drive the growth of smart cities into the next decade (Goldsmith and Crawford 2014).

Cross-References

- [Internet of Things \(IoT\)](#)
- [Open Data](#)

Further Reading

- Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, 22(1), 3–21.
- Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3), 274–279.
- Campbell, T. (2013). *Beyond smart cities: How cities network, learn and innovate*. New York: Routledge.
- Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of Urban Technology*, 18(2), 65–82.
- de Lange, M., & de Waal, M. (2013). Owning the city: New media and citizen engagement in urban design. *First Monday*, 18(11). doi:10.5210/fm.v18i11.4954.
- Goldsmith, S., & Crawford, S. (2014). *The responsive city: Engaging communities through data-smart governance*. San Francisco: Jossey-Bass.
- Hollands, R. G. (2008). Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? *City*, 12(3), 303–320.
- Komninos, N. (2015). *The age of intelligent cities: Smart environments and innovation-for-all strategies*. New York: Routledge.
- Meijer, A., & Bolívar, M. P. R. (2015). Governing the smart city: A review of the literature on smart urban governance. *International Review of Administrative Sciences*. doi:10.1177/0020852314564308.
- Sadowski, J., & Pasquale, F. A. (2015). The spectrum of control: A social theory of the smart city. *First Monday*, 20(7). doi:10.5210/fm.v20i7.5903.
- Schrock, A. R. (2016). Civic hacking as data activism and advocacy: A history from publicity to open government data. *New Media & Society*, 18(4), 581–599.
- Townsend, A. (2013). *Smart cities: Big data, civic hackers, and the quest for a new utopia*. New York: W.W. Norton.
- Vanolo, A. (2013). Smartmentality: The smart city as disciplinary strategy. *Urban Studies*, 51(5), 883–898.

Social Media

Dimitra Dimitrakopoulou
School of Journalism and Mass Communication,
Aristotle University of Thessaloniki,
Thessaloniki, Greece

Social media and networks are based on the technological tools and the ideological foundations of Web 2.0 and enable the production, distribution, and exchange of user-generated content. They transform the global media landscape by

transposing the power of information and communication to the public that had until recently a passive role in the mass communication process.

Web 2.0 tools refer to the sites and services that emerged during the early 2000s, such as blogs (e.g., Blogspot, Wordpress), wikis (e.g., Wikipedia), microblogs (e.g., Twitter), social networking sites (e.g., Facebook, LinkedIn), video (e.g., YouTube), image (e.g., Flickr), file-sharing platforms (e.g., We, Dropbox), and related tools that allow participants to create and share their own content. Though the term was originally used to identify the second coming of the Web after the dotcom burst and restore confidence in the industry, it became inherent in the new WWW applications through its widespread use.

The popularity of Web 2.0 applications demonstrates that, regardless of their levels of technical expertise, users can wield technologies in more active ways than had been apparent previously to traditional media producers and technology innovators. In addition to referring to various communication tools and platforms, including social networking sites, social media also hint at a cultural mindset that emerged in the mid-2000s as part of the technical and business phenomenon referred to as Web 2.0.

It is important to distinguish between social media and social networks. Whereas often both terms are used interchangeably, it is important to understand that social media are based on user-generated content produced by the active users who now can act as producers as well. Social media have been defined on multiple levels, starting from more operational definitions that underline that social media indicate a shift from HTML-based linking practices of the open Web to linking and recommendation, which happen inside closed systems. Web 2.0 has three distinguishing features: it is easy to use, it facilitates sociality, and it provides users with free publishing and production platforms that allow them to upload content in any form, be it pictures, videos, or text. Social media are often contrasted to traditional media by highlighting their distinguishing features, as they refer to a set of online tools that supports social interaction between users. The term is often used to contrast

with more traditional media such as television and books that deliver content to mass populations but do not facilitate the creation or sharing of content by users as well as their ability to blur the distinction between personal communication and the broadcast model of messages.

Theoretical Foundations of Social Media

Looking into the role of the new interactive and empowering media, it is important to study their development as techno-social systems, focusing on the dialectic relation of structure and agency. As Fuchs (2014) describes, media are techno-social systems, in which information and communication technologies enable and constrain human activities that create knowledge that is produced, distributed, and consumed with the help of technologies in a dynamic and reflexive process that connects technological structures and human agency. The network infrastructure of the Internet allows multiple and multi-way communication and information flow between agents, combining both interpersonal (one-to-one), mass (one-to-many), and complex, yet dynamically equal communication (many-to-many).

The discussion on the role of social media and networks finds its roots in the emergence of the network society and the evolvement of the Internet as a result of the convergence of the audiovisual, information technology, and telecommunications sector. Contemporary society is characterized by what can be defined as convergence culture (Jenkins 2006) in which old and new media collide, where grassroots and corporate media intersect, where the power of the media producer and the power of the media consumer interact in unpredictable ways.

The work of Manuel Castells (2000) on the network society is central, emphasizing that the dominant functions and processes in the Information Age are increasingly organized around networks. Networks constitute the new social morphology of our societies, and the diffusion of networking logic substantially modifies the operation and outcomes in processes of production,

experience, power, and culture. Castells (2000) introduces the concept of “flows of information,” underlining the crucial role of information flows in networks for the economic and social organization.

In the development of the flows of information, the Internet holds the key role as a catalyst of a novel platform for public discourse and public communication. The Internet consists of both a technological infrastructure and (inter)acting humans, in a technological system and a social subsystem that both have a networked character. Together these parts form a techno-social system. The technological structure is a network that produces and reproduces human actions and social networks and is itself produced and reproduced by such practices.

The specification of the online platforms, such as Web 1.0, Web 2.0, or Web 3.0, marks distinctively the social dynamics that define the evolution of the Internet. Fuchs (2014) provides a comprehensive approach for the three “generations” of the Internet, founding them on the idea of knowledge as a threefold dynamic process of cognition, communication, and cooperation. The (analytical) distinction indicates that all Web 3.0 applications (cooperation) and processes also include aspects of communication and cognition and that all Web 2.0 applications (communication) also include cognition. The distinction is based on the insight of knowledge as threefold process that all communication processes require cognition, but not all cognition processes result in communication, and that all cooperation processes require communication and cognition, but not all cognition and communication processes result in cooperation.

In many definitions, the notions of collaboration and collective actions are central, stressing that social media are tools that increase our ability to share, to cooperate, with one another, and to take collective action, all outside the framework of traditional institutional institutions and organizations. Social media enable users to create their own content and decide on the range of its dissemination through the various available and easily accessible platforms. Social media can serve as online facilitators or enhancers of human

networks – webs of people that promote connectiveness as a social value.

Social network sites (SNS) are built on the pattern of online communities of people who are connected and share similar interests and activities. Boyd and Ellison (2007) provide a robust and articulated definition of SNS, describing them as Web-based services that allow individuals to (1) construct a public or semipublic profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site. As the social media and user-generated content phenomena grew, websites focused on media sharing began implementing and integrating SNS features and becoming SNSs themselves.

The emancipatory power of social media is crucial to understand the importance of networking, collaboration, and participation. These concepts, directly linked to social media, are key concepts to understand the real impact and dimensions of contemporary *participatory media culture*. According to Jenkins (2006), the term participatory culture contrasts with older notions of passive media consumption. Rather than talking about media producers and consumers occupying separate roles, we might now see them as participants who interact with each other and contribute actively and prospectively equally to social media production.

Participation is a key concept that addresses the main differences between the traditional (old) media and the social (new) media and focuses mainly on the empowerment of the audience/users of media toward a more active information and communication role. The changes transform the relation between the main actors in political communication, namely, political actors, journalists, and citizens. Social media and networks enable any user to participate in the mediation process by actively searching, sharing, and commenting on available content. The distributed, dynamic, and fluid structure of social media enables them to circumvent professional and political restrictions on news production and

has given rise to new forms of journalism defined as citizen, alternative, or participatory journalism, but also new forms of propaganda and misinformation.

The Emergence of Citizen Journalism

The rise of social media and networks has a direct impact on the types and values of journalism and the structures of the public sphere. The transformation of interactions between political actors, journalists and citizens through the new technologies has created the conditions for the emergence of a distinct form from professional journalism, often called citizen, participatory, or alternative journalism. The terms used to identify the new journalistic practices on the Web range from interactive or online journalism to alternative journalism, participatory journalism, citizen journalism, or public journalism. The level and the form of public's participation in the journalistic process determine whether it is a synergy between journalists and the public or exclusive journalistic activities of the citizens.

However, the phenomenon of alternative journalism is not new. Already in the nineteenth century, the first forms of alternative journalism made their appearance with the development of the radical British press. The radical socialist press in the USA in the early twentieth century followed as did the marginal and feminist press between 1960 and 1970. Fanzines and zines appeared in the 1970s and were succeeded by pirate radio stations. At the end of the twentieth century, however, the attention has moved to new media and Web 2.0 technologies.

The evolution of social networks with the new paradigm shift is currently defining to a great extent the type, the impact, and the dynamics of action, reaction, and interaction of the involved participants in a social network. According to Atton (2003), alternative journalism is an ongoing effort to review and challenge the dominant approaches to journalism. The structure of this alternative journalistic practice appears as the counterbalance to traditional and conventional media production and disrupts its dominant

forms, namely, the institutional dimension of mainstream media, the phenomena of capitalization and commercialization, and the growing concentration of ownership.

Citizen journalism is based on the assumption that the public space is in crisis (institutions, politics, journalism, political parties). It appears as an effort to democratize journalism and thereby is questioning the added value of objectivity, which is supported by professional journalism.

The debate on a counterweight to professional, conventional, mainstream journalism was intensified around 1993, when the signs of fatigue and the loss of public's credibility in journalism became visible and overlapped with the innovative potentials of the new interactive technologies. The term public journalism (public journalism) appeared in the USA in 1993 as part of a movement that expressed concerns for the detachment of journalists and news organizations from the citizens and communities, as well as of US citizens from public life. However, the term citizen journalism has defined on various levels. If both its supporters and critics agree on one core thing, it is that it means different things to different people.

The developments that Web 2.0 has introduced and the subsequent explosive growth of social media and networks mark the third phase of public journalism and its transformation to alternative journalism. The field of information and communication is transformed into a more participatory media ecosystem, which evolves the news as social experiences. News are transformed into a participatory activity to which people contribute their own stories and experiences and their reactions to events.

Citizen journalism proposes a different model of selection and use of sources and of news practices and redefinition of the journalistic values. Atton (2003) traces the conflict with traditional, mainstream journalism in three key points: (a) power does not come exclusively from the official institutional institutions and the professional category of journalists, (b) reliability and validity can derive from descriptions of lived experience and not only objectively detached reporting, and (c) it is not mandatory to separate the facts from

subjective opinion. Although Atton (2003) does not consider lived experiences as an absolute value, he believes it can constitute the added value of alternative journalism, combining it with the capability of recording it through documented reports.

The purpose of citizen journalism is to reverse the "hierarchy of access" as it was identified by Glasgow University Media Group, giving voice to the ones marginalized by the mainstream media. While mainstream media rely extensively on elite groups, alternative media can offer a wider range of "voices" that wait to be heard. The practices of alternative journalism provide "first-hand" evidences, as well as collective and anti-hierarchical forms of organizations and a participatory, radical approach of citizen journalism. This form of journalism is identified by Atton as native reporting.

To determine the moving boundary between news producers and the public, Bruns (2005) used the term *produsers*, combining the words and concepts of producers and users. These changes determine the way in which power relations in the media industry and journalism are changing, shifting the power from journalists to the public.

Social Movements

In the last few years, we have witnessed a growing heated debate among scholars, politicians, and journalists regarding the role of the Internet in contemporary social movements. Social media tools such as Facebook, Twitter, and YouTube which facilitate and support user-generated content have taken up a leading role in the development and coordination of a series of recent social movements, such as the student protests in Britain at the end of 2010 as well as the outbreak of revolution in the Arab world, the so-called Arab Spring.

The open and decentralized character of the Internet has inspired many scholars to envisage a rejuvenation of democracy, focusing on the (latent) democratic potentials of the new media as interactive platforms that can motivate and fulfill the active participation of the citizens in

the political process. On the other hand, Internet skeptics suggest that the Internet will not itself alter traditional politics. On the contrary, it can generate a very fragmented public sphere based on isolated private discussions while the abundance of information, in combination with the vast amounts of offered entertainment and the options for personal socializing, can lead people to restrain from public life. The Internet actually offers a new venue for information provision to the citizen-consumer. At the same time, it allows politicians to establish direct communication with the citizens free from the norms and structural constraints of traditional journalism.

Social media aspire to create new opportunities for social movements. Web 2.0 platforms allow protestors to collaborate so that they can quickly organize and disseminate a message across the globe. By enabling the fast, easy, and low-cost diffusion of protest ideas, tactics, and strategies, social media and networks allow social movements to overcome problems historically associated with collective mobilization.

Over the last years, the center of attention was not the Western societies, which were used in being the technology literate and information-rich part of the world, but the Middle Eastern ones. Especially after 2009, there is considerable evidence advocating in favor of the empowering, liberating, and yet engaging potentials of the online social media and networks as in the case of the protesters in Iran who have actively used Web services like Facebook, Twitter, Flickr, and YouTube to organize, attract support, and share information about street protests after the June 2009 presidential elections. More recently, a revolutionary wave of demonstrations has swept the Arab countries as the so-called Arab Spring, using again the social media as means for raising awareness, communication, and organization, facing at the same time strong Internet censorship. Though neglecting the complexity of these transformations, the uprisings were largely quoted as “the Facebook revolution,” demonstrating the power of networks.

In the European continent, we have witnessed the recent development of the Indignant Citizens Movement, whose origin was largely attributed to

the social movements that started in Spain and then spread to Portugal, the Netherlands, the UK, and Greece. In these cases, the digital social networks have proved powerful means to convey demands for a radical renewal of politics based on a stronger and more direct role of citizens and on a critique of the functioning of Western democratic systems.

Cross-References

- [Digital Literacy](#)
- [Open Data](#)
- [Social Network Analysis](#)

Further Reading

- Atton, C. (2003). What is ‘alternative’ journalism? *Journalism: Theory Practice and Criticism*, 4(3), 267–272.
- Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.
- Bruns, A. (2005). *Gatewatching: Collaborative online news production*. New York: Peter Lang.
- Castells, M. (2000). *The rise of the network society, the information age: Economy, society and culture vol. I*. Oxford: Blackwell.
- Fuchs, C. (2014). *Social media: A critical introduction*. London: Sage.
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: New York University Press.

Social Media and Security

Samer Al-khateeb¹ and Nitin Agarwal²

¹Creighton University, Omaha, NE, USA

²University of Arkansas Little Rock, Little Rock, AR, USA

Introduction

In a relatively short period of time, online social networks (OSNs) such as Twitter, Facebook, YouTube, and blogs have revolutionized how societies interact. While this new phenomenon in

online socialization has brought the world closer, OSNs have also led to new vectors to facilitate cybercrime, cyberterrorism, cyberwarfare, and other deviant behaviors perpetrated by state/non-state actors (Agarwal et al. 2017; Agarwal and Bandeli 2018; Galeano et al. 2018; Al-khateeb and Agarwal 2019c).

Since OSNs are continuously producing data with heightened volume, variety, veracity, and velocity, traditional methods of forensic investigation would be insufficient, as this data would be real time, constantly expanding, and simply not found in traditional sources of forensic evidence (Huber et al. 2011; Al-khateeb et al. 2016). These newer forms of data, such as the communications of hacker groups on OSNs, would offer insights into, for example, coordination and planning (Al-khateeb et al. 2016, 2018). Social media is growing as a data source for cyber forensics, providing new types of artifacts that can be relevant to investigations (Baggili and Breitingner 2015). Al-khateeb and Agarwal (2019c) identified key social media data types (e.g., text posts, friends/groups, images, geolocation data, demographic information, videos, dates/times), as well as their corresponding applications to cyber forensics (author attribution, social network identification, facial/object recognition, personality profiling, location finding, cyber-profiling, deception detection, event reconstruction, etc.). Practitioners must embrace the idea of using real-time intelligence to assist in cyber forensic investigations, and not just postmortem data.

Due to afforded anonymity and perceived less personal risk of connecting and acting online, deviant groups are becoming increasingly common among socio-technically competent “hactivist” groups to provoke hysteria, coordinate (cyber) attacks, or even effect civil conflicts. Such deviant groups are categorized as the new face of transnational crime organizations (TCOs) that could pose significant risks to social, political, and economic stability. Online deviant groups have grown in parallel with OSNs, whether it is:

- Black hat hackers who use Twitter to recruit and arm attackers, announce operational details, coordinate cyberattacks (Al-khateeb et al. 2016), and

post instructional or recruitment videos on YouTube targeting certain demographics

- State/non-state actors’ and extremist groups’ (such as ISIS’) savvy use of social communication platforms to make their message viral by using social bots (Al-khateeb and Agarwal 2015c)
- Conduct phishing operations, such as viral retweeting a message containing image which if clicked unleashes malware (Calabresi 2017)

The threat these deviant groups pose is real and can manifest in several forms of deviance, such as the disabling of critical infrastructure (e.g., the Ukraine power outage caused by Russian-sponsored hackers that coordinated a cyberattack in December 2015) (Volz and Finkle 2016). All this necessitate expanding the traditional definitions of cyber threats from hardware attacks and malware infections to include such insidious threats that influence behaviors and actions, using social engineering and influence operations (Carley et al. 2018). Observable malicious behaviors in OSNs, similar to the aforementioned ones, continue to negatively impact society warranting their scientific inquiry. It would benefit information assurance (IA) domain, and its respective subdomains, to conduct novel research on the phenomenon of deviant behavior in OSNs and especially the communications on social platforms pertaining to the online deviant groups.

Definitions

Below are some of the terms that are used in topics related to social media and security and also are frequently used in this entry. *Online deviant groups (ODGs)* refer to groups of individuals that are connected online using social media platforms or the dark web and have interest in conducting deviant acts or events (e.g., disseminating false information, hacking). These events or acts are unusual, unaccepted, and illegal and can have significant harmful effects on the society and public in general. ODGs conduct their activities for various financial or ideological purposes because these ODGs could include state and non-

state actors, e.g., the so-called Islamic State in Iraq and Levant (ISIL), anti-NATO propagandist (Al-khateeb et al. 2016), Deviant Hackers Networks (DHNs) (Al-khateeb et al. 2016), and Internet trolls (Sindelar 2014). ODGs can conduct various deviant acts such as Deviant Cyber Flash Mobs (DCFM); online propaganda, misinformation, or disinformation dissemination; and recently deepfake.

Flash mob (FM) is a form of public engagement, which, according to Oxford dictionaries, is defined as “a large public gathering at which people perform an unusual or seemingly random act and then quickly disperse” (Oxford-Dictionary 2004). Recent observations pertaining to the deviant aspect of the flash mobs have insisted to add a highly debated perspective, which is the nature of the flash mob, i.e., whether it is for entertainment, satire, and artistic expression (e.g., group of people gather and dance in a shopping mall), or it is a deviant act that can lead to robberies and thefts such as the “bash mob” that happened in Long Beach, California in July 9, 2013 (Holbrook 2013). Deviant Cyber Flash Mobs (DCFM) are defined as the cyber manifestation of flash mobs (FM). They are known to be coordinated via social media, telecommunication devices, or emails and have a harmful effect on one or many entities such as government(s), organization(s), society(ies), and country(ies). These DCFMs can affect the physical space, cyberspace, or both, i.e., the “cybernetic space” (Al-khateeb and Agarwal 2015b).

Organized propaganda, misinformation, or disinformation campaigns by group of individuals using social media, e.g., Twitter, are considered as an instance of a DCFM (Al-khateeb and Agarwal 2015a). For example, the dissemination of ISIL’s beheading video-based propaganda of the Egyptians Copts in Libya (Staff 2015), the Arab-Israeli “Spy” in Syria (editorial 2015), and the Ethiopian Christians in Libya (Shaheen 2015). ISIL’s Internet recruitment propaganda or the E-Jihad is very effective in attracting new group members (News 2014). For example, a study conducted by Quiggle (2015) on the effects of developing high production value beheading videos and releasing on social media by ISIL members shows that ISIL’s disseminators are

excellent narrators and they choose their symbols very carefully to give the members of the groups the feeling of pride as well as cohesion. The beheading of civilians has been studied in the literature by Regina Janes (2005). In her study, Janes categorized the reasons for why beheading is done into four main categories, viz., judicial, sacrificial, presentational, and trophy. ISIL’s communicators designed the beheading videos to serve all of the four categories.

In addition to the aforementioned acts, ODGs are increasingly disseminating deepfake videos. Deepfake is defined as a technology that uses a specific type of artificial intelligence algorithms called “generative adversarial network” to alter or produce fake videos. This technique has been used in the past, probably by technical savvy hobbyists to create pornographic videos of various celebrities; however, in many recent cases, these videos targeted more political figures such as the current president Donald Trump, ex-president Barack Obama, Nancy Pelosi, etc. The sophistication and ease of use of these algorithms gave the capability to anyone to produce high-quality deepfaked videos that are nearly impossible for the humans or machines to distinguish from the real one. This type of videos are very dangerous as they can mislead citizen to believe in various lies (imagine a deepfaked video showing a president of a specific nation saying that they just launched a nuclear attack on another nation! If this video is taken seriously by the adversary nation, it can lead to a war or an international catastrophe) and also can lead citizen to distrust real videos (Purdue 2019; “What is deepfake (deep fake AI)?” 2019).

Many of the ODGs conduct their deviant activities using deviant actors, who can be real human (e.g., Internet trolls) or nonhuman (e.g., social bots). Internet trolls are deviant groups who flourished as the Internet become more social, i.e., with the advent of social media. These groups disseminate provocative posts on social media for the troll’s amusement or financial incentives (Davis 2009; Indiana University 2013; Moreau n.d.). Their provocative posts, e.g., insulting a specific person or group, posting false information, or propaganda on popular social media sites, result in a flood of angry responses and often

hijack the discussion (Davis 2009; Moreau *n.d.*; Sindelar 2014). Such “troll armies” (or “web brigades”) piggyback on the popularity of social media to disseminate fake pictures and videos coordinating effective disinformation campaigns to which even legitimate news organizations sometimes fall prey (Sindelar 2014).

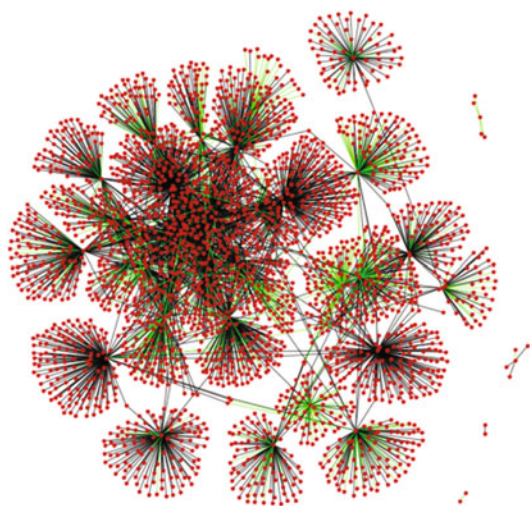
In addition to the human actors, ODGs use nonhuman actors such as social bots to conduct their deviant acts. Social bots are computer programs that can be designed and scheduled to perform various tasks on behalf of the bot’s owner or creator. Research shows that most Internet traffic, especially on social media, is generated by bots (Cheng and Evans 2009). Bots can have a benign intention such as the Woebot which is a chatbot that helps people track their mood and give them therapeutic advices; however, it can also have a malicious intent such as hacker bots, spambots, and deviant social bots which are designed to act like a human in social spaces, e.g., social media, and can influence people’s opinion by disseminating propaganda, disinformation, etc. (@botnerds 2017).

State of the Art

Very little scientific treatment has been given to the topic of social cyber forensics (Carley et al. 2018; Al-khateeb and Agarwal 2019b). Most cyber security research in this direction to date (e.g., Al Mutawa et al. 2012; Mulazzani et al. 2012; Walnycky et al. 2015) has focused on the acquisition of social data from digital devices and the applications installed on them. However, this data would be based on the analysis of the more traditional sources of evidence found on systems and devices, such as file systems and captured network traffic. Since online social networks (OSNs) are continuously creating and storing data on multiple servers across the Internet, traditional methods of forensic investigation would be insufficient (Huber et al. 2011). As OSNs continuously replace traditional means of digital storage, sharing, and communication (Galeano et al. 2019), collecting this ever-growing volume of data is becoming a challenge. Within the past

decade, data collected from OSNs has already played a major role as evidence in criminal cases, either as incriminating evidence or to confirm alibis. Interestingly, despite the growing importance of data that can be extracted from OSNs, there has been little academic research aimed at developing and enhancing techniques to effectively collect and analyze this data (Baggili and Breitingner 2015).

In this entry the aim is to take steps toward bridging the gap between cyber security, big data analytics, and social computing. For instance, in one such study, Al-khateeb et al. (2016) collected Twitter communications network of known hacker groups and analyzed their messages and network for several weeks (Fig. 1). After applying advanced text analysis and social network analysis techniques, it was observed that hacktivist groups @OpAnonDown and @CypherLulz communicate together a lot more than the rest of the nodes. Similarly, members of the “think tank” group and the “cult of the dead cow” group are very powerful/effective in coordination strategies. Furthermore, these groups use Twitter highly effectively to spread their messages via hashtags such as #TangoDown (indicating a successful attack), #OpNimr (calling for DDOS attacks on Saudi Arabian Government websites), #OpBeast



Social Media and Security, Fig. 1 Communication network of black hat hacker accounts on Twitter

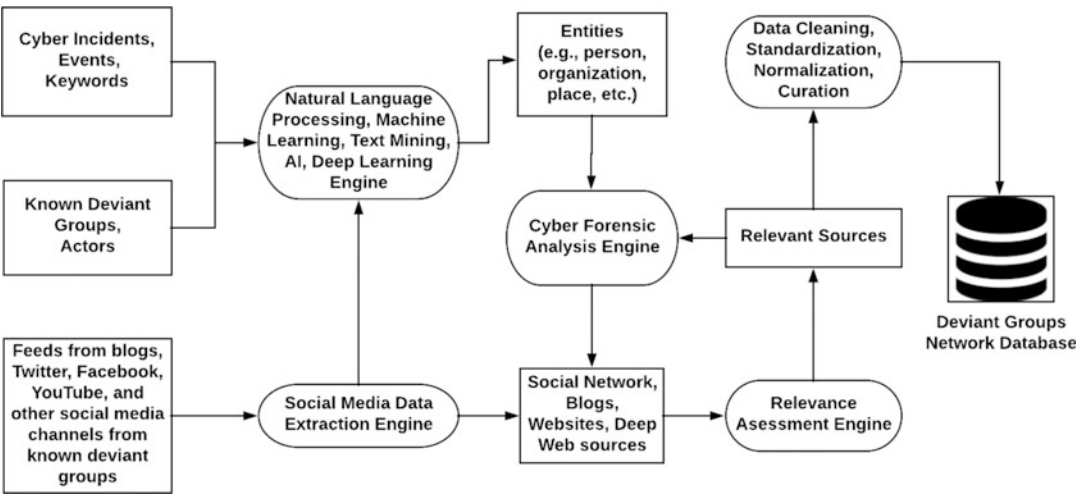
(calling for DDOS attacks on animal rights groups’ websites), among others. Based on work conducted by some of the researchers in this domain, it is clear that OSNs contain vast amounts of important and often publicly accessible data that can service cyber forensics and related disciplines. A progression must thus be made toward developing and/or adopting methodologies to effectively collect and analyze evidentiary data extracted from OSNs and leverage them in relevant domains outside of classical information sciences.

Research Methods

To accomplish the aforementioned research thrusts, socio-computational models are developed to advance our understanding of online deviant groups (ODGs) networks (Al-khateeb 2017; Al-khateeb and Agarwal 2019a, b) grounded in the dynamics of various social and communication processes such as group formation, activation, decentralized decision-making, and collective action. Leveraging cyber forensics and deep web search-based methodologies, the study extracts relevant open-source information in a guided snowball data collection manner. Further,

existing research helps in identifying key actors (Agarwal et al. 2012) and key groups (Sen et al. 2016) responsible for coordinating cyberattacks. At a more fundamental level, embracing the theories of collective action and collective identity formation, the research identify the necessary conditions that lead to the success or failure of coordinated cyberattacks (e.g., phishing campaigns), explain the risk and motivation trade-off governing the sustenance of such coordinated acts, and develop predictive models of ODGs.

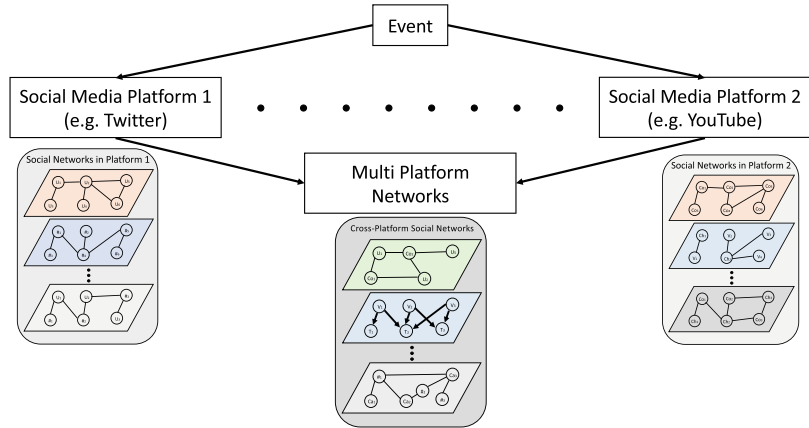
The methodology can be separated into two main phases: (1) data acquisition and (2) data analysis and model development. The main tasks of phase 1 (Fig. 2) include identifying keywords, events, and cyber incidents, selecting reliable and optimal social media sources, and collecting the information from relevant social media sources and metadata using social cyber forensics. The data is largely unstructured and noisy and that warrants cleaning, standardization, normalization, and curation before proceeding to phase 2. Phase 2 entails categorizing cyber incidents, analyzing incident reports with geolocations to identify geospatial diffusion patterns, correlating the identified incidents to current news articles, and examining ODGs’ social and communication networks to identify prominent actors and groups, their



Social Media and Security, Fig. 2 Social media data collection and curation methodology

Social Media and Security, Fig. 3

Multilayer network analysis of deviant groups' social media communications



tactics, techniques, procedures (TTPs), and coordination strategies. A multilayered network analysis approach (Fig. 3) is adopted to model multisource, supra-dyadic relations, and shared affiliations among DGNs.

Research Challenges

Prominent challenges that the research is confronted with include unstructured data, data quality issues/noisy social media data, privacy, and data ethics. Collecting data from social media platforms poses certain limitations including sample bias, missing and noisy information along with other data quality issues, data collection restrictions, and privacy-related issues. Agarwal and Liu (2009) present existing state-of-the-art solutions to the problems mentioned above, in particular, unstructured data mining, noise filtering, and data collection from social media platforms. Due to privacy concerns, data collection can only be observational in nature. Furthermore, only publicly available information can be collected, and all personally identifiable information needs to be stripped before publishing the data, results, or studies. Collecting cyber incidents from social media is prone to sample bias due to the inherent demographic bias among the social media users. The research needs to evaluate this bias by comparing the social media

accounts with mainstream media reports on longitudinal basis and developing corrective measures for reliable analysis.

Conclusion

In conclusion, social media as good as it is in connecting people around the globe, forming communities and partnerships, getting customer feedback on various products, getting real-time news updates, marketing, and advertising, it also poses a security risk on the society, intrudes privacy, helps in radicalizing citizens, provokes hysteria, and provides a fertile ground for ODGs to conduct various deviant events. In the era of big data and exascale computing (the fastest supercomputer in the world), the government, industry, academic, and the public should work together to dismantle any risk that can be posed by social media. Although there are currently various initiatives that are trying to address the many security risks posed by social media, e.g., the HONR Network which is a company that helps the families of people who are affected by mass shootings deal with possible misinformation aftermath (Thompson 2019) and the Facebook “Deepfake Detection Challenge” which aims at creating deepfake videos that can be used by the artificial intelligence (AI) research community to help their algorithms detect fake

videos (Metz 2019), more efforts should be invested.

Further Reading

Social media and security are an inherently multi-disciplinary and multi-methodological computational social science. Researchers in this area employ multi-technology computational social science tool chains (Benigni and Carley 2016) employing network analysis and visualization (Carley et al. 2016), language technologies (Hu and Liu 2012), data mining and statistics (Agarwal et al. 2012a), spatial analytics (Cervone et al. 2016), and machine learning (Wei et al. 2016). The theoretical results and analytics are often multilevel focusing simultaneously on change at the community and conversation level, change at the individual and group level, and so forth.

Acknowledgments This research is funded in part by the US National Science Foundation (OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), US Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17-1-2675, N00014-19-1-2336), US Air Force Research Lab, US Army Research Office (W911NF-16-1-0189), US Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, and the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

- @botnerds. (2017). Types of bots: An overview of chatbot diversity|botnerds.com. Retrieved September 7, 2019, from Botnerds website: <http://botnerds.com/types-of-bots/>.
- Agarwal, N., & Bandeli, K. (2018). Examining strategic integration of social media platforms in disinformation campaign coordination. *Journal of NATO Defence Strategic Communications*, 4, 173–206.
- Agarwal, N., & Liu, H. (2009). *Modeling and data mining in blogosphere*. San Rafael, California (USA): Morgan & Claypool.
- Agarwal, N., Liu, H., Tang, L., & Yu, P. (2012a). Modeling blogger influence in a community. *Social Network Analysis and Mining*, 2(2), 139–162. Springer.
- Agarwal, N., Kumar, S., Gao, H., Zafarani, R., & Liu, H. (2012b). Analyzing behavior of the influentials across social media. In L. Cao & P. Yu (Eds.), *Behavior computing* (pp. 3–19). London: Springer. https://doi.org/10.1007/978-1-4471-2969-1_1.
- Agarwal, N., Al-khateeb, S., Galeano, R., & Goolsby, R. (2017). Examining the use of botnets and their evolution in propaganda dissemination. *Journal of NATO Defence Strategic Communications*, 2, 87–112.
- Al Mutawa, N., Baggili, I., & Marrington, A. (2012). Forensic analysis of social networking applications on mobile devices. *Digital Investigation*, 9, S24–S33.
- Al-khateeb, S. (2017). *Studying online deviant groups (ODGs): A socio-technical approach leveraging social network analysis (SNA) & social cyber forensics (SCF) techniques – ProQuest*. Ph.D. dissertation, University of Arkansas at Little Rock. Retrieved from <https://search.proquest.com/openview/fd4ee2e2719ccf1327e03749bf450a96/1?pq-origsite=gscholar&cbl=18750&diss=y>.
- Al-khateeb, S., & Agarwal, N. (2015a). Analyzing deviant cyber flash mobs of ISIL on Twitter. In *Social computing, behavioral-cultural modeling, and prediction* (pp. 251–257). UCDC Center, Washington DC, USA: Springer.
- Al-khateeb, S., & Agarwal, N. (2015b). Analyzing flash mobs in cybernetic space and the imminent security threats a collective action based theoretical perspective on emerging sociotechnical behaviors. In *2015 AAAI spring symposium series*. Palo Alto, California: Association for the Advancement of Artificial Intelligence.
- Al-khateeb, S., & Agarwal, N. (2015c). *Examining botnet behaviors for propaganda dissemination: A case study of ISIL's beheading videos-based propaganda* (pp. 51–57). Atlantic City, New Jersey, USA: IEEE.
- Al-khateeb, S., & Agarwal, N. (2019a). Deviance in social media. In S. Al-khateeb & N. Agarwal (Eds.), *Deviance in social media and social cyber forensics: Uncovering hidden relations using open source information (OSINF)* (pp. 1–26). Cham: Springer. https://doi.org/10.1007/978-3-030-13690-1_1.
- Al-khateeb, S., & Agarwal, N. (2019b). *Deviance in social media and social cyber forensics: Uncovering hidden relations using open source information (OSINF)*. Cham: Springer.
- Al-khateeb, S., & Agarwal, N. (2019c). Social cyber forensics: Leveraging open source information and social network analysis to advance cyber security informatics. *Computational and Mathematical Organization Theory*. <https://doi.org/10.1007/s10588-019-09296-3>.
- Al-khateeb, S., Conlan, K. J., Agarwal, N., Baggili, I., & Breiting, F. (2016). Exploring deviant hacker

- networks (DHN) on social media platforms. *The Journal of Digital Forensics, Security and Law: JDFSL*, 11(2), 7–20.
- Al-khateeb, S., Hussain, M. N., & Agarwal, N. (2017a). Social cyber forensics approach to study Twitter's and blogs' influence on propaganda campaigns. In D. Lee, Y.-R. Lin, N. Osgood, & R. Thomson (Eds.), *Social, cultural, and behavioral modeling* (pp. 108–113). Washington D.C., USA: Springer International Publishing.
- Al-khateeb, S., Hussain, M., & Agarwal, N. (2017b). Chapter 12: Analyzing deviant socio-technical behaviors using social network analysis and cyber forensics-based methodologies. In O. Savas & J. Deng (Eds.), *Big data analytics in cybersecurity and its management*. New York: CRC Press, Taylor & Francis.
- Al-khateeb, S., Hussain, M., & Agarwal, N. (2018). Chapter 2: Leveraging social network analysis & cyber forensics approaches to study cyber propaganda campaigns. In T. Ozyer, S. Bakshi, & R. Alhaji (Eds.), *Social network and surveillance for society* (Lecture notes in social networks) (pp. 19–42). Springer International Publishing AG, part of Springer Nature: Springer.
- Baggili, I., & Breiting, F. (2015). Data sources for advancing cyber forensics: What the social world has to offer. In *2015 AAAI spring symposium series*. Stanford University, CA.
- Benigni, M., & Carley, K. M. (2016). From tweets to intelligence: Understanding the Islamic jihad supporting community on Twitter. In K. Xu, D. Reitter, D. Lee, & N. Osgood (Eds.), *SBP-BRIMS 2016* (Lecture notes in computer science) (Vol. 9708, pp. 346–355). Cham: Springer. https://doi.org/10.1007/978-3-319-39931-7_33.
- Calabresi, M. (2017). Inside Russia's social media war on America. *Time*. <http://time.com/4783932/inside-russia-social-media-war-america/>. Last accessed 26 Dec 2018.
- Carley, K. M., Wei, W., & Joseph, K. (2016). High dimensional network analytics: Mapping topic networks in Twitter data during the Arab spring. In S. Cui, A. Hero, Z.-Q. Luo, & J. Moura (Eds.), *Big data over networks*. Boston: Cambridge University Press.
- Carley, K., Cervone, G., Agarwal, N., & Liu, H. (2018). *Social cyber-security. International conference on social computing, behavioral-cultural modeling and prediction (SBP-BRIMS)*, July 10–July 13, Washington, DC, USA, pp. 389–394.
- Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J., & Waters, N. (2016). Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. *International Journal of Remote Sensing*, 37(1), 100–124.
- Cheng, A., & Evans, M. (2009). Inside Twitter an in-depth look at the 5% of most active users. Retrieved from Sysomos website: <http://sysomos.com/insidetwitter/mostactiveusers>.
- Davis, Z. (2009, March 24). Definition of: Trolling [Encyclopedia]. Retrieved April 4, 2017, from PCMag.COM website: <http://www.pcmag.com/encyclopedia/term/53181/trolling#>.
- editorial, T. news. (2015). ISIL executes an Israeli Arab after accusing him of been an Israeli spy. *TV7 Israel News*. <http://www.tv7israelnews.com/isil-executes-an-israeli-arab-after-accusing-him-of-been-an-israeli-spy/>. Last checked: June 11, 2015.
- Galeano, R., Galeano, K., Al-khateeb, S., Agarwal, N., & Turner, J. (2018). Chapter 10: Botnet evolution during modern day large scale combat operations. In C. M. Vertuli (Ed.), *Large scale combat operations: Information operations: Perceptions are reality*. Army University Press.
- Galeano, K., Galeano, R., Al-khateeb, S., & Agarwal, N. (2019). Studying the weaponization of social media: A social network analysis and cyber forensics informed exploration of disinformation campaigns. In *Open source intelligence and security informatics*. Springer. (forthcoming).
- Holbrook, B. (2013). LBPd prepared for potential bash mob event. In *Everything Long Beach*. <http://www.everythinglongbeach.com/lbpd-prepared-for-potential-bash-mob-event/>. Last checked: August 15, 2014.
- Hu, X., & Liu, H. (2012). Text analytics in social media. In C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 385–414). Boston: Springer. https://doi.org/10.1007/978-1-4614-3223-4_12.
- Huber, M., Mulazzani, M., Leithner, M., Schrittwieser, S., Wondracek, G., & Weippl, E. (2011). Social snapshots: Digital forensics for online social networks. In *Proceedings of the 27th annual computer security applications conference* (pp. 113–122). Orlando, Florida, USA
- Indiana University. (2013, January 3). What is a troll? [University Information Technology Services]. Retrieved April 4, 2017, from Indiana University Knowledge Base website: <https://kb.iu.edu/d/afhc>.
- Janes, R. (2005). *Losing our heads: Beheadings in literature and culture*. NYU Press.
- Metz, R. (2019, September 5). Facebook is making deepfake videos to help fight them [CNN]. Retrieved September 7, 2019, from <https://www.cnn.com/2019/09/05/tech/facebook-deepfake-detection-challenge/index.html>.
- Moreau, E. (n.d.). Here's what you need to know about internet trolling. Retrieved February 6, 2018, from Lifewire website: <https://www.lifewire.com/what-is-internet-trolling-3485891>.
- Mulazzani, M., Huber, M., & Weippl, E. (2012). Social network forensics: Tapping the data pool of social networks. In *Eighth annual IFIP WG* (Vol. 11). University of Pretoria, Pretoria, South Africa: Springer
- News, C. B. S. (2014). ISIS recruits fighters through powerful online campaign. <http://www.cbsnews.com/news/>

- [isis-uses-social-media-to-recruit-western-allies/](#). Last checked: July 1, 2015.
- Oxford-Dictionary. (2004). Definition of flash mob from Oxford English Dictionaries Online. In *Oxford English Dictionaries*. <http://www.oxforddictionaries.com/definition/english/flash-mob>. Last checked: August 22, 2014.
- Purdue, M. (2019, August 14). Deepfake 2020: New artificial intelligence is battling altered videos before elections. Retrieved September 6, 2019, from USA TODAY website: <https://www.usatoday.com/story/tech/news/2019/08/14/election-2020-company-campaigns-against-political-deepfake-videos/2001940001/>.
- Quiggle, D. (2015). The ISIS beheading narrative. *Small Wars Journal*. Retrieved from <https://smallwarsjournal.com/jrnl/art/the-isis-beheading-narrative>.
- Sen, F., Wigand, R., Agarwal, N., Yuce, S., & Kasprzyk, R. (2016). Focal structures analysis: Identifying influential sets of individuals in a social network. *Journal of Social Network Analysis and Mining*, 6(1), 1–22. Springer.
- Shaheen, K. (2015). Isis video purports to show massacre of two groups of Ethiopian Christians. *The Guardian*. <http://www.theguardian.com/world/2015/apr/19/isis-video-purports-to-show-massacre-of-two-groups-of-ethiopian-christians>. Last checked: June 11, 2015.
- Sindelar, D. (2014). The Kremlin's troll Army: Moscow is financing legions of pro-Russia internet commenters. But how much do they matter? *The Atlantic*. Retrieved from <http://www.theatlantic.com/international/archive/2014/08/the-kremlins-troll-army/375932/>.
- Staff, C. (2015, February 16). ISIS video appears to show beheadings of Egyptian Coptic Christians in Libya [News Website]. Retrieved January 23, 2017, from CNN website: <http://www.cnn.com/2015/02/15/middleeast/isis-video-beheadings-christians/>.
- Thompson, N. (2019, July 10). A Grieving Sandy Hook Father on How to Fight Online Hoaxers. Retrieved September 7, 2019, from Medium website: <https://onezero.medium.com/a-grieving-sandy-hook-father-on-how-to-fight-online-hoaxers-ce2e0ef374c3>.
- Volz, D. and Finkle, J. (2016) U.S. helping Ukraine investigate power grid hack. Reuters. January 12. <https://www.reuters.com/article/us-ukraine-cybersecurity-usaidUSKCN0UQ24020160112>. Last accessed 26 Dec 2018.
- Walnycky, D., Baggili, I., Marrington, A., Moore, J., & Breitingner, F. (2015). Network and device forensic analysis of android social-messaging applications. *Digital Investigation*, 14, S77–S84.
- Wei, W., Joseph, K., Liu, H., & Carley, K. M. (2016). Exploring characteristics of suspended users and network stability on Twitter. *Social Network Analysis and Mining*, 6(1), 51.
- What is deepfake (deep fake AI)? – Definition from WhatIs.com. (2019). Retrieved September 6, 2019, from WhatIs.com website: <https://whatistechtarget.com/definition/deepfake>.

Social Network Analysis

Magdalena Bielenia-Grajewska

Division of Maritime Economy, Department of Maritime Transport and Seaborne Trade, University of Gdansk, Gdansk, Poland
Intercultural Communication and Neurolinguistics Laboratory, Department of Translation Studies, University of Gdansk, Gdansk, Poland

Social Network Analysis: Origin and Introduction

The origins of Social Network Theory can be observed in the works of such sociologists as Ferdinand Tönnies, Émile Durkheim, and Georg Simmel, as well as in the works devoted to sociometry, such as the one by Jacob Moreno on sociograms. Moreover, researchers interested in holism study the importance of structure over individual entities and the way structures govern the performance of people. Although the interest on social networks can be traced back to the previous centuries, its great popularity can be observed in modern times. The reasons for this state are as follows. First of all, technology has led to the proliferation of social networks, nowadays also available on the web. Thus, an individual has the opportunity to enter into relationships not only in the “standard” way, but also in the online one, by participating in online discussion lists or social online networking tools. In addition, the performance of social networks in the offline mode is supported by the advancements of technology; an example can be the use of mobile telephones to stay in contact with other network members. Secondly, technological advancements have led to the emergence of data that require a proper methodological approach, taking into account the perspective of human beings as both authors and subjects of a given big data study. Thirdly, individuals are more and more conscious about the significance of social networks in their life. Starting from the

microlevel, people are embedded in different networks, such as families, communities, or professional groups. Looking at the issue from the macrolevel, nowadays the world can be viewed as a complex system, being made of different networks, of both national and international character, that concern, among others, such areas of life as economics, transportation, energy, as well as private and social life. The definition of a social network provided by Wasserman and Faust (1994) in Devan, Barnett and Kim (2011: 27) is as follows: *a social network is generally defined as a system with a set of social actors and a collection of social relations that specify how these actors are relationally tied together.*

Social Networks: Main Determinants and Characteristics

Technology belongs to the main factors of modern social networks since it offers the creation of new types of social networks and supports the ones existing in the offline sphere. Taking into account modern technological advancements, the Internet constitutes an important factor responsible for creating and sustaining the growth in the area of social networks. The developments in the sphere of online communication have led to the proliferation of social contacts existing on the web. Social networking tools are used for both professional and private purposes, being the place where an individual meets with friends and family as well as with the ones he or she does not know at all. Taking into account the mentioned online dimension of social networks, these networks serve different purposes. Their functionality can be viewed from the perspective of synchronicity. Synchronous social networks require the real-time participation in discussion, whereas asynchronous social networks do not require immediate response by users. As far as the synchronous and asynchronous social networks are concerned, they are used, among others, to talk (e.g., Skype), share photos or videos (e.g., Picassa and YouTube), connect with friends, acquaintance, or customers (e.g., Facebook), or search for professional

opportunities (e.g., LinkedIn). Another important factor for shaping social networks is language. Individuals select the social networks that offer them the possibility to interact in the language they know. It should be mentioned, however, that the relation between social networks and language is mutual. Language does not only shape social networks, being the tool of efficient communication among network members. At the same time, social networks create linguistic repertoires, being visible in new terms and forms of expressions created during the interaction among network members. An example can be the language used by the users of discussion lists who coin new terms to denote the reality around them. Another important factor for creating and sustaining social network is information. Networks are crucial because of the growing role of data, innovation, and knowledge and new possibilities of creating and disseminating knowledge; only the ones who have access to information and can distribute it effectively can compete on the modern market. Thus, information is linked with competition visible from both organizational and individual perspectives. Starting with the organizational dimension, companies that cooperate with others in terms of purchasing raw materials, production, and distribution have a chance to be successful on the competitive market. It should be stated, however, that competition is not exclusively the feature of business entities since individuals also have to be competitive, e.g., on the job market. The interest in continuous education has led to the popularization of open universities or online courses, and, consequently, new social networks formed within massive open online courses (MOOCs) have been created. The next determinant for forming social networks that should be stressed is the need for belonging and socializing. Individuals need the contact with others to share their feelings and emotions, to have fun and to quarrel. In the case of those who have to be far away from their relatives and friends, online social networks have become the sphere of socialization and interaction.

The mentioned multifactorial and multi-aspectual character of social networks has resulted

in the intense studies on methodologies underlying the way social networks are formed and exercised as well as their role for the environment in the micro, meso, and macro meaning.

Main Concepts and Terms in Social Network Analysis (SNA)

Social Network Analysis can be defined as an approach that aims to study how the systems of grids, ties and lattices create human relations. Social Network Analysis focuses on both internal and external features shaping social networks, studying individuals or organizations, and the relations between them. Social networks are also studied by taking into account intercultural differences. Applying the determinants used to characterize national or professional cultures, researchers may study how social networks are formed and organized by taking into account the attitude to hierarchy, punctuality, social norms, family values, etc. Social Network Analysis is mainly used in behavioral and social studies, but it is also applied in different disciplines, such as marketing, economics, linguistics, management, biology, neuroscience, cognitive studies, etc. SNA relies on the following terminology, with many terms coming from the graph theory. As Wasserman and Faust (1994) stress, the fundamental concepts in SNA are: actor, relational tie, dyad, triad, subgroup, and group. *Actors* (or vertices, nodes) are social entities, such as individuals, companies, groups of people, communities, nation states, etc. An example of an actor is a student at the university or a company operating in one's neighborhood. *Ego* is used to denote a focal actor. The characteristics of an actor are called *actor attributes*. *Relational ties* constitute the next important notions, being the linkages used to transfer material and immaterial resources, such as information, knowledge, emotions, products, etc. They include one's personal feelings and opinions on other people or things (like, dislike, hatred, love), contacts connected with the change of ownership (purchasing and selling, giving and receiving things) or changes of geographical, social, or professional position (e.g., becoming

expatriate, marrying a person of a higher social status, receiving a promotion). Relational ties are determined by such factors as geographical environments and interior designs. Taking the example of corporations, such notions as the division of office space or the arrangement of managerial offices reflects the creation of social networks. Relational ties may also be shaped by, e.g., time differences that determine the possibility of participation in online networks. Relational ties may also be governed by the type of access to communication tools, such as mobile telephones, social networking tools and the Internet. Relational ties may be influenced by other types of networks, such as transport or economic networks. In addition, relational ties are connected with the flows of ideas and things as well as the movement of people. For example, expatriates form social networks in host countries. Taking the reason into consideration, relational ties may be formed for private and professional reasons. As far as the private domain is concerned, relational ties are connected with one's need for emotional support, intimacy, or sharing common hobbies. Relational ties are formed in a voluntary and involuntary way. *Voluntary relational ties* are connected with one's free will to become the members of a group or close friendship with another person. On the other hand, *involuntary relational ties* may be of biological origin (family ties) or hierarchical notions, such as relations at work. The number of actors participating in a given social network can be analyzed through the dyad or triad perspective. Dyad involves two actors and their relations, whereas triads concern three actors and their relations. Another classification of networks includes groups and subgroups. The set of relational ties constitutes relations. Another term discussed in social actor network theory is the notion of structural holes. Degenne and Forsé (1999) elaborate on the concept of *structural holes* introduced by Burt who states that the structural hole is connected with non-redundancy between contacts. This concept is studied by these two scholars through the prism of cohesion and equivalence. According to the cohesive perspective presented by them, redundancy can be observed when two of the egos' relations have a direct link.

Consequently, when the cohesion is great, few structural holes can be observed. They state that the approach of equivalence is connected with indirect relations in networks between the ego and others. Structural holes exist when there are no direct or indirect links between the ego and the contacts or when there is no structural equivalence between them.

Types of Social Networks

Social networks can be categorized by taking into account different dimensions. One of them is network size; social networks vary as far as the number of network members is concerned. The second notion is the place where the social network is created and exercised. The main dichotomy is of technological nature; the distinction between online and offline social networks is one of the most researched types. The next feature that can subcategorize social networks is formality. *Informal social networks* are mainly used to socialize and entertain, whereas *formal social networks* encompass the contacts characterized by strict codes of behavior in a network, hierarchical relations among network members, and regulated methods of interaction. The next feature of social networks is uniformity. As Bielenia-Grajewska and Gunstone (2015) discuss, *heterogeneous social networks* encompass members that differ as far as certain characteristics are concerned. On the other hand, *homogeneous social networks* include similar network members. The types of member compatibility differ, depending on the characteristics of social networks, and may be connected with, e.g., profession, age, gender, mother tongue, and hobby. In SNA terminology, these networks are often described through the prism of homophily. *Homophilous social networks* consist of people who are similar because of age, gender, social status, cultural background, profession, or hobbies. On the other hand, *heterophilous social networks* are directed mainly at individuals that differ as far as their individual attributes, social or professional positions are concerned. Bielenia-Grajewska (2014) also stresses that networks can also be divided into

horizontal and vertical networks. *Vertical social network* encompass the relations between people that occupy different positions in, e.g., hierarchical ladders. They include networks to be observed in professional and occupational settings, such as organizations, universities, schools, etc. On the other hand, *horizontal social networks* encompass members of an equal position in a given organization. Networks may also be classified by taking into account their power and the strength of relations between networks members. *Weak social networks* are the ones that are loosely composed, with fragile and loose relations between members, whereas in *strong social networks* the contacts are very durable. Social networks can be classified by taking into account the purpose why they were formed. For example, *financial social networks* concern the money-related flows between members, whereas *informational social networks* focus on exchanging information. Networks may also be studied by taking into account their flexibility. Thus, *fixed social networks* rely on a strict arrangement, whereas *flexible social networks* do not follow a fixed pattern of interactions. As far as advantages of social networks are concerned, such notions as the access to information, creating social relations can be named. As far as potential disadvantages are concerned, some state that networks demand the resign from independence. In addition, to some extent the members of a network bear responsibility for the mistakes made by other members since a single failure may influence the performance of the whole network. Analyzing more complex entity networks, they may demand more energy and time to adjust to new conditions. Social Networks can be divided into online social networks and offline social networks. As far as online social networks are concerned, they can be further subcategorized into asynchronous online social networks and synchronous social networks (discussion on these networks provided above). Social networks can be studied through the prism of purpose and the prism of investigation may focus on the dichotomy of private and professional life. For example, *professional social networks* may be categorized by taking into account the notion of free will in network creation and performance. Professional social networks

depend on the type of organizations. For example, at universities student and scientific networks can be examined. They mainly include the relations formed at work or connected with work, as well as the ones among specialists from the same discipline. *Private social networks* are formed and sustained in one's free time to foster family relations, participate in hobby, etc. The next dichotomy may involve the notion of law and order. For example, Social Network Analysis also studies illegal networks and the issue of crime in social networks. Another classification involves the notion of entry conditions. *Closed social networks* are aimed exclusively at carefully selected individuals, providing barriers for entering them. Examples of closed social networks are social networks at work; the reason for their closeness is connected with the need for privacy and openness among the network users. On the contrary, *open social networks* do not pose any entrance barriers for the users. Social networks may also be categorized by taking into account their scope. One of the ways to look at social networks is to divide them into local and global social networks. Depending on other notions, the local character of social networks is connected with the limited scope of its performance, being restricted, e.g., to the local community. Wasserman and Faust (1994) categorize networks by taking modes into account. The mode is understood as the number of sets of entities. *One-mode networks* encompass a single set of entities, whereas *two-mode networks* involve two sets of entities or one set of entities and one set of events. More complex networks are studied through the perspective of three or more mode networks.

Methods of Investigating Social Networks

There are different ways of researching social networks, depending what features of social networks are to be investigated. For example, a researcher may study the influence of social networks on individuals or the types of social networks and their implications for professional or private life. Participants may also be asked to run

a diary and note down the names of people they meet or interact with in some settings. The methods that involve the direct participation of researchers include, e.g., observation and ethnographic studies. Thus, individuals and the relations between them are observed, e.g., in their natural environment, such as at home or at work. Another way of using SNA is by conducting interviews and surveys, asking respondents to answer questions related to their social networks. As in the case of other social research, methods can be divided into qualitative and quantitative ones. Social Network Analysis, as other methods used in social studies, may benefit from neuroscientific investigation, using such techniques as, e.g., fMRI or EEG to study emotions and involvement in communities or groups. Taking the growing role of the Internet, social networks are also analyzed by studying the interactions in the online settings. Thus, SNA concerns the relations taking places in social online networking tools, discussion forums, emails, etc. Social Network Analysis takes into account differences within the online places of interaction, by observing the access to the online tool, types of individuals the tool is directed at, etc. It should be stressed that since social networks do not exist in a vacuum, they should be studied by taking their environment into account. Thus, other network approaches may prove useful to study the relation between different elements in systems. In addition, social network analysis studies not only human beings and organizations and the way they cooperate but also technological elements are taken into account. For example, Actor-Network-Theory (ANT) may prove useful in the discussion on SNA since it facilitates the understanding of the relations between living and non-living entities in shaping social relations. For example, computer network or telephone networks and their influence on social networks may be studied. Moreover, since social networks are often created and exercised in communication, modern methodological approaches include discourse studies, such as Critical Discourse Analysis, that stress how the selection of verbal and nonverbal elements of communication (e.g., drawings, pictures)

facilitates the creation of social networks and their performance.

Social Network Analysis and Big Data Studies

There are different ways social networks are linked with big data. First of all, social networks generate a large amount of data. Depending on the type of networks, big data concern pictorial data, verbal data or audio data. Big data gathered from social networks can also be categorized by taking into account the type of network and accumulated data. For example, professional social networks provide data on users' education and professional experience, whereas private social networks offer information on one's hobbies and interests. Depending on the type of network they are gathered from, data provide information on demographic changes, customer preference and behaviors, etc. One of the ways is to organize information on social groups and professional communities. Social Network Analysis may be applied to the study on modern organizations to show how big data is gathered and distributed in organizations. SNA is also important when there is an outbreak of a disease or other crisis situations to show how information is administered within a given social network. It should be stressed, however, that the process of gathering and storing data should reflect the ethical principles of research. In the case of showing big data, SNA visualization techniques (e.g., VISIONE) facilitate the presentation of complex data. SNA may also benefit from statistics, by applying, e.g., exponential random graph models or such programs as UCINET or PAJEK. Big data in social networks may be handled in different ways, but one of the key problems in such analyses includes memory and time limits. Stanimirović and Mišković (2013) have developed three metaheuristic methods to overcome the mentioned difficulties: a pure evolutionary algorithm (EA), a hybridization of the EA and a Local Search Method (EA-LS), and a hybridization of the EA and a

Tabu Search heuristic (EA-TS). In addition, the application of SNA in the studies on big data can be analyzed by taking into account different concepts crucial for the research conducted in various disciplines. One of such concept is identity that can be investigated in, e.g., organizational setting. Company identity being understood as the image created at both the external and internal level of corporations is a complex concept that requires multilevel studies. Within the phenomenon of company identity, its linguistic dimension can be studied, by taking into account how communication is created and conducted within corporate social networks. It should also be stated that the study on complex social networks is connected with some problems that researchers may encounter; for example, companies may have different hierarchies and the ways they are organized. One of them is the issue of boundary setting and group membership. It should also be stated that Social Network Analysis relies on different visualization techniques that offer the pictorial presentation of gathered data.

Cross-References

- [Blogs](#)
- [Digital Storytelling, Big Data Storytelling](#)
- [Economics](#)
- [Facebook](#)
- [Network Analytics](#)
- [Network Data](#)
- [Social Media](#)

Further Reading

- Bielenia-Grajewska, M. (2014). Topology of social networks. In K. Harvey (Ed.), *Encyclopedia of social media and politics*. Thousand Oaks: SAGE.
- Bielenia-Grajewska, M., & Gunstone, R. (2015). Language and learning science. In R. Gunstone (Ed.), *Encyclopedia of science education*. Dordrecht: Springer.
- Degenne, A., & Forsé, M. (1999). *Introducing social networks*. London: SAGE.

- Rosen, D., Barnett, G. A., & Kim, J. H. (2011). Social networks and online environments: when science and practice co-evolve. *SOCNET 1*, 27–42. <https://doi.org/10.1007/s13278-010-0011-7>.
- Stanimirović, Z., & Mišković, S. (2013). Efficient meta-heuristic approaches for exploration of online social networks. In W.-C. Hu & N. Kaabouch (Eds.), *Data management, technologies, and applications*. Hershey: IGI Global.
- Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge: Cambridge University Press.

Social Sciences

Ines Amaral
University of Minho, Braga, Minho, Portugal
Instituto Superior Miguel Torga, Coimbra,
Portugal
Autonomous University of Lisbon, Lisbon,
Portugal

Social Science is an academic discipline concerned with the study of humans through their relations with society and culture. Social Science disciplines analyze the origins, development, organization, and operation of human societies and cultures. The technological evolution has strengthened Social Sciences since it enables empirical studies developed through quantitative means, allowing the scientific reinforcement of many theories about the behavior of man as a social actor. The rise of big data represents an opportunity for the Social Sciences to advance the understanding of human behavior using massive sets of data.

The issues related to Social Sciences began to have a scientific nature in the eighteenth century with the first studies on the actions of humans in society and their relationships with each other. It was by this time that Political Economy emerged. Most of the subjects belonging to the fields of Social Sciences, such as Anthropology, Sociology, and Political Science arisen in the nineteenth century.

Social Sciences can be divided in disciplines that are dedicated to the study of the evolution of

societies (Archeology, History, Demography), social interaction (Political Economy, Sociology, Anthropology), or cognitive system (Psychology, Linguistics). There are also applied Social Sciences (Law, Pedagogy) and other Social Sciences classified in the generic group of Humanities (Political Science, Philosophy, Semiotics, Communication Sciences). The anthropologist Claude Lévi-Strauss, the philosopher and political scientist Antonio Gramsci, the philosopher Michel Foucault, the economist and philosopher Adam Smith, the economist John Maynard Keynes, the psychoanalyst Sigmund Freud, the sociologist Émile Durkheim, the political scientist and sociologist Max Weber, and the philosopher, sociologist, and economist Karl Marx are some of the leading social scientists of the last centuries.

The social scientist studies phenomena, structures, and relationships that characterize the social and cultural organizations; analyzes the movements and population conflicts, the construction of identities, and the formation of opinions; researches behaviors and habits and the relationship between individuals, families, groups, and institutions; and develops and uses a wide range of techniques and research methods to study human collectivities and understand the problems of society, politics, and culture.

The study of humans through their relations with society and culture relied on “surface data” and “deep data.” “Surface data” was used in the disciplines that adapted quantitative methods, like Economics. “Deep data” about individuals or small groups was used in disciplines that analyze society through qualitative methods, such Sociology.

Data collection has always been a problem for social research because of its inherent subjectivity as Social Sciences have traditionally relied on small samples using methods and tools gathering information based on people. In fact, one of the critical issues of Social Science is the need to develop research methods that ensure the objectivity of the results. Moreover, the objects of study of Social Sciences do not fit into the models and methods used by other sciences and do not allow the performance of experiments under controlled

laboratory conditions. The quantification of information is possible because there are several techniques of analysis that transform ideas, social capital, relationships, and other variables from social systems into numerical data. However, the object of study always interacts with the culture of the social scientist, making it very difficult to have a real impartiality.

Big data is not self-explanatory. Consequently, it requires new research paradigms across multiple disciplines, and for social scientists, it is a major challenge as it enables interdisciplinary studies and the intersection between computer science, statistics, data visualization, and social sciences. Furthermore, big data empowers the use real-time data on the level of whole populations, to test new hypotheses and study social phenomena on a larger scale. In the context of modern Social Sciences, large datasets allow scientists to understand and study different social phenomena, from the interactions of individuals and the emergence of self-organized global movements to political decisions and the reactions of economic markets.

Nowadays, social scientists have more information on interaction and communication patterns than ever. The computational tools allow understanding the meaning of what those patterns reveal. The models build about social systems within the analysis of large volumes of data must be coherent with the theories of human actors and their behavior. The advantages of large datasets and of the scaling up the size of data are that it is possible to make sense of the temporal and spatial dimensions. What makes big data so interesting to Social Sciences is the possibility to reduce data, apply filters that allow to identify relevant patterns of information, aggregate sets in a way that helps identify temporal scales and spatial resolutions, and segregate streams and variables in order to analyze social systems.

As big data is dynamic, heterogeneous, and interrelated, social scientists are facing new challenges due to the existence of computational and statistical tools, which allow extracting and analyzing large datasets of social information. Big data is being generated in multiple and

interconnecting disciplinary fields. Within the social domain, data is being collected from transactions and interactions through multiple devices and digital networks. The analysis of large datasets is not within the field of a single scientific discipline or approach. In this regard, big data can change Social Science because it requires an intersection of sciences within different research traditions and a convergence of methodologies and techniques. The scale of the data and the methods required to analyze them need to be developed combining expertise with scholars from other scientific disciplines. Within this collaboration with data scientists, social scientists must have an essential role in order to read the data and understand the social reality.

The era of big data implies that Social Sciences rethink and update theories and theoretical questions such as small world phenomenon, complexity of urban life, relational life, social networks, study of communication and public opinion formation, collective effervescence, and social influence. Although computerized databases are not new, the emergence of an era of big data is critical as it creates a radical shift of paradigm in social research. Big data reframes key issues on the foundation of knowledge, the processes and techniques of research, the nature of information, and the classification of social reality.

The new forms of social data have interesting dimensions: volume, variety, velocity, exhaustive, indexical, relational, flexible, and scalable. Big data consists of relational information in large scale that can be created in or near real time with different structures, extensive in scope, capable of identifying and indexing information distinctively, flexible, and able to expand in size quickly.

The datasets can be created by personal data or nonpersonal data. Personal data can be defined as information relating to an identified person. This definition includes online user-generated content, online social data, online behavioral data, location data, sociodemographic data, and information from an official source (e.g., police records). All data collected that do not directly identify individuals are considered nonpersonal data. Personal data can be collected from different sources with

three techniques: voluntary data that is created and shared online by individuals; observed data, which records the actions of the individual; and data inferred about individuals based on voluntary information or observed.

The disciplinary outlines of Social Sciences in the age of big data are in constant readjustment because of the speed of change in the data landscape. Some authors argued that the new data streams could reconfigure and constitute social relations and populations. Academic researchers attempt to handle the methodological challenges presented by the growth of big social data, and new scientific trends arise, although the diversity of the philosophical foundations of Social Science disciplines. Objectivity of the data does not result directly in their interpretation. The scientific method postulated by Durkheim attempts to remove itself from the subjective domain. Nevertheless, the author stated that objectivity is made by subjects and is based on subjective observations and selections of individuals.

A new empiricist epistemology emerged in Social Sciences and goes against the deductive approach that is hegemonic within modern science. According to this new epistemology, big data can capture an entire social reality and provide their full understanding. Therefore, there is no need for theoretical models or hypotheses. This perspective assumes that patterns and relationships within big data are characteristically significant and accurate. Thus, the application of data analytics transcends the context of a single scientific discipline or a specific domain of knowledge and can be interpreted by those who can interpret statistics or data visualization.

Several scholars, who believe that the new empiricism operates as a discursive rhetorical device, criticize this approach. Kitchin argues that whereas data can be interpreted free of context and domain-specific expertise, such an epistemological interpretation is probable to be unconstructive as it absences to be embedded in broader discussions.

As large datasets are highly distributed and present complex data, a new model of data-driven science is emerging within the Social Science

disciplines. The data-driven science uses a hybrid combination of abductive, inductive, and deductive methods to the understanding of a phenomenon. This approach assumes theoretical frameworks and pursues to generate scientific hypotheses from the data by incorporating a mode of induction into the research design. Therefore, the epistemological strategy adopted within this model is to detect techniques to identify potential problems and questions, which can be worth of further analysis, testing, and validation.

Although big data enhance the set of data available for analysis and enable new approaches and techniques, it does not replace the traditional small data studies. Due to the fact that big data cannot answer specific social questions, more targeted studies are required. Computational Social Sciences can be the interface between computer science and the traditional social sciences. This interdisciplinary and emerging scientific from Social Sciences uses computationally methods to model social reality and analyze phenomena, as well as social structures and collective behavior. The main computational approaches from Social Sciences to study big data are social network analysis, automated information extraction systems, social geographic information systems, complexity modeling, and social simulation models.

Computational Social Science is an intersection of Computer Science, Statistics, and the Social Sciences, which uses large-scale demographic, behavioral, and network data to analyze individual activity, collective behaviors, and relationships. Computational Social Sciences can be the methodological approach to Social Sciences study big data because of the use of mathematical methods to model social phenomena and the ability to handle with large datasets.

The analysis of big volumes of data opens up new perspectives of research and makes it possible to answer questions that were previously incomprehensible. Though big data itself is relative, its analysis within the theoretical tradition of Social Sciences to build a context for information will enable its understanding and the intersection with the smaller studies to explain specific data variables.

Big data may have a transformational impact as it can transform policy making, by helping to improve communication and governance in several policy domains. Big social data also raise significant ethical issues for academic research and request an urgent debate for a wider critical reflection on the epistemological implications of data analytics.

Cross-References

- [Anthropology](#)
- [Communications](#)
- [Complex Networks](#)
- [Computational Social Sciences](#)
- [Computer Science](#)
- [Data Science](#)
- [Network Analytics](#)
- [Network Data](#)
- [Psychology](#)
- [Social Network Analysis](#)
- [Visualization](#)

Further Reading

- Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, 55(1), 193–196.
- Berg, B. L., & Lune, H. (2004). *Qualitative research methods for the social sciences* (Vol. 5). Boston: Pearson.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.
- Coleman, J. S. (1990). *Foundations of social theory*. Cambridge, MA: Belknap Press of Harvard University Press.
- Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, 25, 435–437.
- González-Bailón, S. (2013). Social science in the era of big data. *Polymer International*, 5(2), 147–160.
- Lohr, S. (2012). The age of big data. *New York Times* 11.
- Lynch, C. (2008). Big data: How do your data grow? *Nature*, 455(7209), 28–29.
- Oboler, A., et al. (2012). The danger of big data: Social media as computational social science. *First Monday*, 17(7-2). Retrieved from <http://firstmonday.org/ojs/index.php/fm/article/view/3993/3269>.

Socio-spatial Analytics

Xinyue Ye

Landscape Architecture & Urban Planning, Texas A&M University, College Station, TX, USA

Questions on inequality lie at the heart of the discipline of social science and geography, motivating the development of socio-spatial analytics. Growing socioeconomic inequality across various spatial scales in any region threatens social harmony. Meanwhile, a number of fascinating debates on the trajectories and mechanisms of socioeconomic development are reflected in numerous empirical studies ranging from specific regions and countries to the scale of individuals and groups. With accelerated technological advancements and convergence, there have been major changes in how people carry out their activities and how they interact with each other. With these changes in both technology and human behavior, it is imperative to improve our understanding of human dynamics in order to tackle the inequality challenges ranging from climate change, public health, traffic congestion, economic growth, digital divide, social equity, political movements, and cultural conflicts, among others. Socio-spatial analytics has been, and continues to be, challenged by dealing with the temporal trend of spatial patterns and spatial dynamics of social development from the human needs perspective. As a framework promoting human-centered convergence research, socio-spatial analytics has the potential to enable more effective and symbiotic collaboration across disciplines to improve human societies. The growth in citizen science and smart cities has reemphasized the importance of socio-spatial analytics. Theory, methodology, and practice of computational social science have emerged as an active domain to address these challenges.

Socio-spatial analytics can reveal the dynamics of spatial economic structures, such as the emergence and evolution of poverty traps and convergence clubs. Spatial inequality is multiscale in

nature. Sources or underlying forces of inequality are also specific to geographic scale and social groups. Such a scalar perspective presents a topology of inequality and has the potential to link inequalities at the macroscale to the microscale, even everyday life experiences. The dramatic improvement in computer technology and the availability of large-volume geographically referenced social data have enabled spatial analytical methods to move from the fringes to central positions of methodological domains. The history of the open-source movement is much younger, but its impact on quantitative social science and spatial analysis is impressive. The OSGeo projects that support spatial data handling have a large developer community with extensive collaborative activities, possibly due to the wide audience and publicly adopted OGC standards. In comparison, spatial analysis can be quite flexible and is often field- and data-specific. Therefore, analysis routines are often written by domain scientists with specific scientific questions in mind. The explosion of these routines is also facilitated by increasingly easier development processes with powerful scripting language environments such as R and Python.

In addition to space, things near in time or in statistical distribution are more related than distant things. Hence, ignoring the interdependence across space, time, and statistical distribution leads to overlooking many possible interactions and dependencies among space, time, and attributes. To reveal these relationships, the distributions of space, time, and attributes should be treated as the context in which a socio-spatial measurement is made, instead of specifying a single space or time as the context. The “distribution” in space (the dimension of space) refers to the spatial distribution of attributes, while the “distribution” of attributes (the dimension of statistical distribution) implies the arrangement of attributes showing their observed or theoretical frequency of occurrence. In addition, the “distribution” of time (the dimension of time) signifies the temporal trend of attributes.

To advance core knowledge on how humans understand and communicate spatial relationships

under different contexts, many efforts have aimed at analyzing fine-scale spatial patterns and geographical dynamics and maximizing the potential of massive data to improve human well-being toward human dynamics level. Big spatiotemporal data have become increasingly available, allowing the possibility for individuals’ behavior in space and time to be modeled and for the results of such models to be used to gain information about trends at a daily and street scale. Many research efforts in socioeconomic inequality dynamics can be substantially transformed in the context of new data and big data. Spatial inequality can be further examined at the finer scale, such as social media data and movement data, in order to catalyze knowledge and action on environment and sustainability challenges in the built environment. Given the multidimensionality, current research faces challenges of systematically uncovering spatiotemporal and societal implications of human dynamics. Particularly, a data-driven policy-making process may need to use data from various sources with varying resolutions, analyze data at different levels, and compare the results with different scenarios. As such, a synthesis of varying spatiotemporal and network methods is needed to provide researchers and planning specialists a foundation for studying complex social and spatial processes. Socio-spatial analytics can be delivered in an interactive visual system to answer day-to-day questions by non-specialized users. The following questions can be asked: has this policy change brought any positive effects to this street? Where can I tell my patient to exercise that is safe, culturally acceptable, and appropriate to who he/she is?

Professionals working with academic, government, industry, and not-for-project organizations across the socio-spatial analytics also recognize a widespread challenge with adequately implementing spatial decision support capabilities to address complex sustainable systems problems. Wide-ranging knowledge gaps stem in large part from an inability to synthesize data, information, and knowledge emerging from diverse stakeholder perspectives broadly, deeply, and flexibly within application domains of spatial decision support systems. Socio-spatial

analytics will facilitate an understanding of the complicated mechanisms of human communications and policy development in both cyberspace (online) and the real world (offline) for decision support.

The metrics for evaluation for socio-spatial analytics can cover the following quantitative and qualitative aspects: (1) usability – whether the proposed solutions achieve the goal of understanding the socioeconomic dynamics; (2) acceptability, whether and to what degree the proposed solutions are operational for dissemination into the community; (3) extensibility, whether the proposed methodology and workflow can be used to address different themes; (4) documentation, whether the documentation has sufficient and clear descriptions about the proposed solutions as well as software tools; and (5) community building, whether and how this research can attract the attention and participation of researchers from diverse domain communities and how the research can be extended to other themes.

Cross-References

- ▶ [Social Sciences](#)
- ▶ [Spatiotemporal Analytics](#)

Further Reading

- Ye, X., & He, C. (2016). The new data landscape for regional and urban analysis. *GeoJournal*. <https://doi.org/10.1007/s10708-016-9737-8>.
- Ye, X., & Mansury, Y. (2016). Behavior-driven agent-based models of spatial systems. *Annals of Regional Science*. <https://doi.org/10.1007/s00168-016-0792-3>.
- Ye, X., & Rey, S. (2013). A framework for exploratory space-time analysis of economic data. *The Annals of Regional Science*, 50(1), 315–339.
- Ye, X., Huang, Q., & Li, W. (2016). Integrating big social data, computing, and modeling for spatial social science. *Cartography and Geographic Information Science*. <https://doi.org/10.1080/15230406.2016.1212302>.
- Ye, X., Zhao, B., Nguyen, T. H., & Wang, S. (2020). Social media and social awareness. In *Manual of digital earth* (pp. 425–440). Singapore: Springer.

South Korea

Jooyeon Lee

Hankuk University of Foreign Studies, Seoul, Korea (Republic of)

Keeping pace with global trends, the Korean government is seeking the use of big data in various areas through efforts such as the publication of the report *A Plan to Implement a Smart Government Using Big Data* by the President's Council on Information Strategies in 2011. In particular, the current government aims for the realization of a smart government that creates convergence of knowledge through data sharing between government departments. One of the practical strategies for achieving these goals is active support of the use of big data in the public sector. For this purpose, the Big Data Strategy Forum was launched in April 2012 led by the National Information Society Agency. In addition to this, the Electronics and Telecommunications Research Institute (ETRI) in South Korea has carried out a task to build up knowledge assets for the use of big data in the public sector with the support of the Korean Communications Commission and the Korean Communications Agency. The South Korean government is also operating a national big data center. The main purpose of this center is to support small- and medium-sized businesses, universities, and institutions that find it difficult to manage or maintain big data due to financial constraints. Furthermore, this center is preparing to develop new business models by collecting data from telecommunications companies, medical services, and property developers.

There are many examples of how much effort the Korean government is making in applying big data in many different public sector organizations. For example, there is a night bus project run by the Seoul Metropolitan Government which was started in response to a night bus service problem in Seoul in 2013. In order to achieve this, the Seoul Metropolitan Government took advantage of aggregated datasets (comprised of around three billion calls and the analysis results of five million

customers who got in and out of cabs) to create a night bus route map. A second example is the National Health Care Service, which is operated by Korean National Health Insurance. It set up services such as Google Flu Trends, which is a web service which estimates influenza activity in South Korea by analyzing big data. The Korean National Health Insurance Service investigates how many people search for the symptoms of a cold, such as a high fever and coughing, on SNS (Social Network Services) and Twitter. In addition, the Ministry of Employment and Labour in South Korea has used big data by consulting records of customer service centers and search engine systems on SNS to predict supply and demand for job prospects in South Korea. Finally, the Ministry of Gender Equality and Family in South Korea has analyzed big data by consulting records about teenagers who are being bullied at school and feel suicidal urges, blogs, and SNS in order to prevent potential teenager delinquency, suicide, disappearance from home, and academic interruption.

In addition, customized marketing strategies based on big data are becoming popular in the commercial sector. This is reflected, for example, in the marketing strategies of credit card companies. Shinhan Card, one of the major credit companies in South Korea, developed a card product known as Code Nine by analyzing the consumption patterns and characteristics of its 22 million customers. In addition, Shinhan opened its Big Data Center in December 2013. Similarly, Samsung Card is also inviting experts to assist in promoting the use of big data in its business and has opened a marketing-related department responsible for big data analysis. There have been many other attempts at using big data to analyze sociopolitical phenomena, such as public opinion on sensitive political issues.

However, despite such high interest in big data, the big data market in South Korea is still smaller than in other developed countries, such as the United States and the United Kingdom. Moreover, although South Korea is a leader in the IT industry and has the highest Long-Term Evolution (LTE) distribution rate among all Asian countries,

its big data market is over 2 years behind the Chinese equivalent and does not have sufficient specialists with practical skills. In addition, there are problems in that big data has yet not been fully utilized and has not even been discussed fully in South Korea. Even though many scholars in social sciences have stressed the importance of the practical use of big data, it is true that there have been many problems in using big data in reality. Furthermore, there have been many debates about the leakage of personal information resulting from the use of big data, as in other countries. In 2012, credit card companies experienced considerable issues due to a series of leakages of customer information and, as a result, investments and development related to big data have shrunk markedly. Thus, the Korean Communications Commission has been establishing and promoting guidelines for the protection of personal information in big data. The main purposes of the guidelines are to prevent the misuse/abuse of personal information and to delimit the scope of personal information collectible and usable without the information subjects' prior consent within the current legal provisions on the big data industry. Although many civic organizations are involved in arguments for and against the establishment of the guidelines, the Korean government's efforts in the use of big data continue.

Only a few years ago, big data was merely an abstract concept in South Korea. Now, however, the customized services and marketing strategies of Korean companies using extensive personal information are emerging as crucial. In response to this trend, government departments such as the Ministry of Science and Technology and the National Information Society Agency are supporting the big data industry actively; for example, the announcement of the Manual for Big Data Work Processes and Technologies 1.0 on May 2014 for the introduction and distribution of big data services in South Korea. Moreover, from 2015, national qualification examinations have been introduced and academic departments relating to big data will be opened in universities in order to develop big data experts for the future. Furthermore, many social scientists have published many articles related to the practical

use of big data and have discussed how the Korean government and companies will be able to fully use big data and problems they may face. Thus, it is clear that the big data industry in South Korea has been rapidly developed and the efforts of the Korean government and businesses alike in using big data are sure to continue.

Cross-References

- [Data Mining](#)
- [Google Flu](#)
- [Industrial and Commercial Bank of China](#)

Further Reading

- Diana, M. (2014, June 24). Could big data become big brother? *Government Health IT*.
- Jee, K., & Kim, G. H. (2013). Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system. *Healthcare Informatics Research*, 19(2), 79–85.
- Lee, Y., & Chang, H. (2012). Ubiquitous health in Korea: Progress, barriers, and prospects. *Healthcare Informatics Research*, 18(4), 242–251.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Washington, DC: McKinsey Global Institute.
- Park, H. W., & Leydesdorff, L. (2013). Decomposing social and semantic networks in emerging “big data” research. *Journal of Informetrics*, 7(3), 756–765.

Space Research Paradigm

Hina Kazmi

George Mason University, Fairfax, VA, USA

Post-Space Age Advancements

While the field of astronomy is as old as humanity itself, technologically speaking a large part of the scientific understanding about our own planet, the solar system, the Milky Way galaxy, and beyond has occurred in just past 40–60 years. Between the 1960s and the turn of the new millennia, the

primary focus of space research has been in advancing and maturing capabilities in satellites, telescopes, and auxiliary science instruments. These advancements have provided unprecedented multi-spectral, multi-temporal, and multi-spatial data along with the computing abilities to combine these multi-dimensional datasets to enrich the study of various sub-systems within our cosmos. In the new millennia, the research in the fields of earth sciences and astronomy have been undergoing revolutionary changes, mainly due to technical capabilities to acquire, store, handle and analyze large volumes of scientific data. These advancements are building upon technical innovation that the space-age accelerated, thus transforming and enriching space research. In this entry space research refers to both earth and astronomical sciences, conducted from both space and ground-based observatories.

Space-Based Earth Sciences

The very first artificial satellite, Sputnik 1, was launched into low earth orbit in 1957 triggering the dawn of space age. It provided the first set of space-based scientific data about earth's upper atmosphere. Soon after, the US Government launched the world's first series of meteorological satellites, TIROS 1 through 10. Although these early satellites had low resolution cameras and limited capacity to store and transmit earth images, they quickly demonstrated the national and economic significance of accurately forecasting and understanding weather patterns. The next series of meteorological satellites included very high-resolution instruments with multi-spectral sensors providing earth images ranging from visible light to infrared bands. The application of weather satellite technology rapidly expanded to study land masses and led to the launch of Landsat spacecraft series. These satellites have provided the world an unmatched recorded history of land-forms and respective changes over the past 40 years, such as tropical rainforests, glaciers, coral reefs, fault lines, and environmental impact of human-driven development like agriculture and deforestation (Tatem et al. 2008). Further advancements in spacecraft technologies enabled National Aeronautics and Space Agency (NASA)

to launch its Earth Observation System (EOS) – a series of large coordinated satellites with multiple highly complex science instruments that have been in operation since the 1990s. The EOS science has expanded our understanding of earth’s environmental state such as cloud formation, precipitation, ice sheet mass, and ozone mapping. During this period the US military also commercialized the Global Positioning System (GPS) technology that led to the development of multiple Geographic Information System (GIS) tools. Today our dependency on continuous earth observation data is fully intertwined with many day-to-day functions like traffic management and timely responses to weather-related emergencies.

Ground- and Space-Based Astronomy

Throughout the twentieth century, the field of astronomy made major breakthroughs (deGrijs and Hughes 2007). Among the notable discoveries are the expansion and the rate of acceleration of our universe, the measurement of cosmic background radiation (confirming the big bang theory), and discovery and understanding of black holes. These revolutionary findings were confirmed in large part due to launch of space-based telescopes, starting in the 1960s. There are two key advantages of conducting astronomical observations from space. First, images are much sharper and stable above earth’s atmosphere; second, we can observe the sky across wavelengths on the electromagnetic spectrum that are not detectable by ground-based telescopes. In the post-space age era, one of the most significant contributions to astronomy has come from Hubble Space Telescope (HST) (NAP 2005), the largest space-based observatory that included a range of instruments to observe in wavelengths spanning from ultraviolet to near-infrared. Hubble’s findings have revolutionized our understanding of the cosmos. It has made over 1.4 million observations to date and created a repository of science data that has resulted in over 17,000 peer-reviewed publications over its 30 years of operational life; its H-Index remains the highest among all observatories (257 as of 2017) with well over 738,000 citations. HST is a historic treasure, and its images have been

embedded beyond science into mainstream culture and arts globally. Hubble’s findings have been complemented with a series of other advanced space telescopes observing in X-ray, gamma-ray, and infrared wavelengths, which are measurable only from space.

At the same time, ground-based observatories have also grown in size and complexity thanks to a series of innovations such as: (a) mirror technologies making telescopes lighter in weight and adjustable to correct for physical deformations over time (namely, active and adaptive optics) and (b) detectors that allow for high resolution and wide angle. Even more, the addition of radio astronomy as an established discipline in the field has added to the ever-growing set of multi-spectral astronomical research.

Paradigm Shift

Sixty-three years after the launch of Sputnik 1, the overall field of space research is diverse, complex, and rich in data. At the same time, the digital age has equipped us with super-computing, large data storage, and sophisticated analytical tools. The combination of these factors is steadily leading us to new realms in research that is more and more driven by big data. Civilian space agencies, such as NASA and European Space Agency (ESA), have promoted open data policy for decades and have helped develop large publicly available online archives to encourage data-intensive research that varies in both depth and breadth. The National Academies of Sciences, Engineering, and Medicine (NASEM) has also focused on the increasing significance of archival data and interestingly has referred to science centers as “archival centers” for scientific research (NAP 2007). Government continues to invest in various data mining tools for the ease of access along with data recipes to facilitate the use of multiple layers of datasets as well as the metadata associated with the processed results, such as catalogs of object positions and brightness.

In 2013, the Public Broadcast Service’s NOVA made a documentary “Earth From Space” using

NASA's earth observation data. Combining the images from multiple satellites, the documentary creatively demonstrated the delicate interconnectedness and interdependence of biological, geological, oceanic, and atmospheric elements of our planet as one ecosystem. Earth scientists are able to now pursue such expansive research which is indeed a study in system of systems at a full planetary scale building on pool of scientific knowledge of past decades that continues to grow.

In astronomy, the age of big data is changing the very nature of conducting astronomical research, and we are in the midst of this paradigm shift. For example, rather than astronomers generating data for their research by proposing to telescopes to observe selected targets in the sky, big data is influencing scientists to instead create their research programs from the large swaths of archival data already available. More and more now, astronomers do not have to spend time observing with a telescope as their predecessors traditionally did. This trend began with Hubble's expansive archives. New observatories like Gaia and Alma are also driving such archival-based research. Gaia is a space-based astrometric telescope that is designed to build a full 3D map of the Milky Way and track the movement of nearly 2 billion stars and other objects in it by scanning the skies multiple times during its rotation around the sun (Castelvecchi 2020). This is an extraordinary scale of information about galactic structure and patterns of movements of stars within it – including our own Sun. In radio astronomy, the ground-based observatory, Alma, is transforming the study of cosmology in a similar manner. Consequently, the next chapters in astrophysics are focusing on finding larger patterns and understanding the structural characteristics of objects in space. Similar to understanding Earth as one system, scientists hope to use big data to ultimately understand the interconnectedness across system of galaxies and stars that can reveal the construct of system of systems at the cosmological scale.

Earth Science Archives

NASA's Earth Science Data System (ESDS) program adds on average 20 TB of new data per day

in its Earth Science Data and Information System (ESDIS 2020). The data volume in 2019 alone totaled to 34 PB, made up of 12,000 datasets, and served 1.3 million users. The ESDS program projects that its total archive volume will reach close to 250 PB by year 2025.

Astronomy Archives

Astronomy is just showing up to the big data world. Civil space agencies are investing resources to develop and mature various calibration algorithms and archives for all astronomical disciplines (planetary, heliophysics, and astrophysics) and across multiple wavelengths on the electromagnetic spectrum. One such data source that is increasing in demand and volume is NASA's Infrared Science Archive (IRSA). It currently holds 1 PB of data from 15 infrared space-based observatories and is in the process of adding a list of ground-based observatories. Hubble Telescope archives contain about 150 TB of data according to NASA's website. The ESA science data center contains about 646 TB of total archival data for all its science missions, and its monthly download rate is about 69 TB (ESDC 2020). The data-intensive observatories (such as Gaia) aim to scan the full sky on repeated basis for multiple years and thus further pushing the big data-driven archival research in astronomy. The following table lists a few examples of such next generation of observatories that are considered game changers due to the volume of data they intend to generate (Table 1).

Analytical Challenges

The devil is in the details of mining, handling, organizing, analyzing, and visualizing ever-growing volumes of archival records. IRSA program cites limitations in computational capabilities and efficiently transferring large volumes of data (Grid n.d.). NASEM emphasizes the need to increase and sustain common archival systems such as IRSA and organizing astronomical archives by wavelengths and using standardized tools that are repeatable and reliable (NAP 2007). NASA, academia, and other stakeholders are partnering to transition to cloud-based technologies where analytical tools are co-located with

Space Research Paradigm, Table 1 Next generation of observatories

Observatory	Data volume	Description
Gaia	1.3 TB with latest data dump	Space-based astrometric telescope building a full 3D map of the Milky Way galaxy and tracking the movement of nearly 2 billion stars and other objects
Alma	Between 200 and 400 PB annually	Ground-based radio observatory that studies the universe
Vera C. Rubin	About 500 PB annually – over 20 TB of data to be processed daily	Ground-based observatory scheduled to start operations in 2023
Square Km Array	600 PB annually	Planned largest radio telescope ever built which will include thousands of dishes and up to a million low-frequency antennas

data archives – therefore resolving the limitation on downloading and migrating large data sizes. For example, Yao et al. propose the Discrete Global Grid System (DGGS) which is a unified spatiotemporal framework combining data storage, analytics, and imaging (Yao et al. 2019) for earth observing data.

Big data analytics are in preliminary phases in space research. It will take an active participation of scientists working in close collaboration with software, IT, and data scientists to develop DGGS-like tools. More importantly, this collaborative approach is needed to incorporate the complicated data calibration and reduction algorithms that scientists can trust. The development of tools that can reliably and efficiently synergize these facets will be necessary to fully realize the potentials and expansion of space research disciplines.

Summary

Space research is evolving in both depth and breadth since the turn of the millennia. The data-driven paradigm shift in research promises to build upon the technological advancements that were made in the post-space age era. Space agencies are investing in analytical tools and building large data archives, turning science centers into archive centers. As a result, the data-intensive transformation in astronomical and earth sciences is truly exciting, and because of it the humanity is at the cusp of understanding the cosmos and our place in it in unprecedented ways and for decades to come.

Cross-References

- Data Brokers and Data Services
- Earth Science

Further Reading

Castelvecchi, D. (2020, December 3). *Best map of Milky Way reveals a billion stars in motion*. Retrieved from Nature <https://www.nature.com/articles/d41586-020-03432-9>.

deGrijs, R., & Hughes, D. W. (2007, October). The top ten astronomical “Breakthroughs” of the 20th century. *The CAP Journal*, 1(1):11–17.

ESDC. (2020). *ESAC science data center*. Retrieved from <https://www.cosmos.esa.int/web/esdc/home>.

ESDIS. (2020, October 20). *Earth science data*. Retrieved from <https://earthdata.nasa.gov/esds/nasa-earth-science-data-systems-program-highlights-2019>.

Grid, O. S. (n.d.). *Astronomy archives are making new science every day*. Retrieved from <https://openscience.grid.org/news/2017/08/29/astronomy-archives.html>.

NAP. (2005). *Assessment of options for extending the life of the Hubble Space Telescope*. National Academies of Sciences. Retrieved from <https://www.nap.edu/read/11169/chapter/5>.

NAP. (2007). *The portals of the universe: The NASA Astronomy Science Centers*. Retrieved from The Portals of the Universe: The NASA Astronomy Science Centers. <https://www.nap.edu/download/11909>.

NASA. (2019, October). *Earth observing systems*. Retrieved from Project Science Office. <https://eosps.nasa.gov/content/nasas-earth-observing-system-project-science-office>.

Tatem, A. J., Goetz, S. J., & Hay, S. I. (2008, September–October). Fifty years of earth-observation satellites. *American Scientist*, 96(5). Retrieved from <https://www.americanscientist.org/article/fifty-years-of-earth-observation-satellites>.

Yao, X., Li, G., Xia, J., Ben, J., Cao, Q., Zhao, L., ... Zhu, D. (2019). Enabling the big earth observation data via cloud computing and DGGS: Opportunities and challenges. *Remote Sensing*, 12(1):62.

Spain

Alberto Luis García

Departamento de Ciencias de la Comunicación Aplicada, Facultad de Ciencias de la información, Universidad Complutense de Madrid, Madrid, Spain

Big Data technology is growing and, in some cases, has already matured, when it is about to be fulfilled 104 years since the publication of MapReduce, the model of massive and distributed computing that marked its beginning as the heart of Hadoop. MapReduce, as defined in IBM webpage, “is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a concrete cluster.” The term MapReduce, as follows in IBM webpage, actually refers to map job, “that takes a set of data and converts it into another set of data where individual elements are broken down into tuples.”

The origins of the use of Big Data in Spain began in local performances in regions such as the Basque Country, Catalonia, and Asturias; however, Spanish public administration has the webpage (<http://datos.gob.es/>) that offers all kinds of public data for reuse in matters private. This website is managed by the Ministry of Industry, Energy and Tourism and the Ministry of Finance and Public Administration, and a key objective is to create a strategy of openness to the management of public sector data for use in business and to promote the necessary transparency of public policies.

In this same line of policy transparency from Big Data management, the government published in April 2014 the II Action Plan for Open Government under the Open Government Partnership, an alliance born in 2011 between 64 countries – including Spain – whose mission is to develop commitments to achieve improvements in the three key aspects of open government: accountability, participation, and transparency. However, these action plans are under public consultation to indicate which important issues

are missing and should be included to achieve further progress in the policy of Open Government of Spain.

At this point, the Spanish government approved Royal Decree 1495/2011, which regulates the tools available in the above described specific website and open data (general, technical, formats, semantics, etc.) and for the case mix of each agency.

The articulation of the data provided by the administration was managed through catalogs. Access allows to introduce to them from a single point to various websites and resources of the central government to provide public information. The data are available, organized, and structured by formats and topics users, among other criteria. Within the website you have the possibility to search for catalogs or applications; there are examples of specific applications like IPlayas, looking for the nearest beaches and coves to your mobile. Therefore, the objective is to provide services that can help you obtain economic returns in strategic sectors of the Spanish economy.

The profile of users is taken into account to interact with the data is of three types:

- **Anonymous** users, i.e., those who can visit all public areas of the site, send suggestions, rate, and comment content.
- Users **infomediaries** who can publish and manage applications in the App Catalog (with prior registration on the portal).
- Users of **public sector**, which are allowed to add and manage their data sets within the data catalog (with prior registration on the portal).

The profile of users undergoing a major influence on the use of Big Data is the infomediary Sector that was defined as set of companies that generate applications, products, and/or added value services for third parties, from public sector information. The Infomediary Sector has been cataloged into subsectors according to the area of reusable information; these areas are: Business/Economy, Geographical/Cartographical, Social–Demographical/Statistical, Legal, Meteorological, Transport,

Information about Museums, Libraries and Cultural Files.

Within the different types of activity would be the most prolific Geographical/Cartographical Information and Business/Economy Information.

The sources of reused information for the activities in the Infomediary Sector are – from the most used until the less used – State Administration, Regional Administration, Local Administration, European Union, Intelligence Agencies, and from another countries. And in the other hand, the Administration itself becomes a client of infomediary companies, but the most important clients of Spanish Infomediarities Companies are Self-Employers Workers, Universities, and in a second level Public Administration and Citizens. The revenue models for payment of services are payment for Works done, Access, Use, Linear subscription, etc, and the products or services offered from the sector are Data Processed, Maps, Raw Data, and Publications. The main Generic Services from Data are Custom Reports, Advices, Comparatives, and Clipping; the main applications are Client Software, Mobile Software, GPS Information, and SMS/Mail Alerts. And the Assessment of Infomediary Sector for Clients is the development for new products and applications and increased customer loyalty.

In Spain, the whole strategy is integrated into the Plan Aporta in which there are three types of users involved in the reuse of information:

- Public bodies or content generators
- Infomediary Sector or generations of applications and value added
- End users are those who use the information

The regulations governing the Big Data in Spain are designed, therefore, in order to deepen the use of two main elements: the reuse of information for business benefit and transparency in governance. This regulation is perfectly integrated in the common strategy of the European Union around Access Info Europe will allow the reuse of information at higher scales access as it relates to public data and geolocation. However, the Spanish government has not yet issued

specific regulations that will allow the implementation of the regulations.

Further Reading

<http://datos.gob.es/>.

<http://datos.gob.es/saber-mas>.

<http://www.access-info.org/es>.

IBM Webpage. (2014). What is map reduce? <http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce>. Accessed Aug 2014.

Spatial Big Data

► Big Geo-data

Spatial Data

Xiaogang Ma

Department of Computer Science, University of Idaho, Moscow, ID, USA

Synonyms

Geographic information; Geospatial data; Geospatial information

Introduction

Spatial property is almost a pervasive component in the big data environment because everything happening on the Earth happens somewhere. Spatial data can be grouped into raster or vector according to the methods used in representations. Web-based services facilitate the publication and use of spatial data legacies, and the crowdsourcing approaches enable people to be both contributors and users of spatial data. Semantic technologies further enable people to link and query the spatial data available on the Web, find patterns of interest, and to use them to tackle scientific and business issues.

Raster and Vector Representations

Spatial data are representations of facts that contain positional values, and geospatial data are spatial data that are about facts happening on the surface of the Earth. Almost everything on the Earth has location properties, so geospatial data and spatial data are regarded as synonyms. Spatial data can be seen almost everywhere in the big data deluge, such as social media data stream, traffic control, environmental sensor monitoring, and supply chain management, etc. Accordingly, there are various applications of spatial data in the actual world. For example, one may find a preferred restaurant based on the grading results on Twitter. A driver may adjust his route based on the real-time local traffic information. An engineer may identify the best locations for new buildings in an area with regular earthquakes. A forest manager may optimize timber production using data of soil and tree species distribution and considering a few constraints such as the requirement of biodiversity and market price.

Spatial data can be divided into two groups: raster representations and vector representations. A raster representation can be regarded as a group of mutually exclusive cells which form the representation of a partition of space. There are two types of raster representations: regular and irregular. The former has cells with same shape and size and the latter with cells of varying shape and size. Raster representations do not store coordinate pairs. In contrast, vector representations use coordinate pairs to explicitly describe a geographic phenomenon. There are several types of vector representations, such as points, lines, areas, and the triangulated irregular networks. A point is a single coordinate pair in a two-dimensional space or a coordinate triplet in a three-dimensional space. A line is defined by two end points and zero to more internal points to define the shape. An area is a partition of space defined by a boundary (Huisman and de By 2009).

The raster representations have simple but less compact data structures. They enable simple implementation of overlays but pose difficulties for the representation of interrelations among

geographic phenomena, as the cell boundaries are independent of feature boundaries. However, the raster representations are efficient for image processing. In contrast, the vector representations have complex data structures but are efficient for representing spatial interrelations. The vector representations work well in scale changes but are hard to implement overlays. Also, they allow the representation of networks and enable easy association with attribute data.

The collection, processing, and output of spatial data are often relevant to a number of platforms and systems, among them the most well-known are the geographic information system, remote sensing, and the global positioning system. A geographic information system is a computerized system that facilitates the phases of data collection, data processing, and data output, especially for spatial data. Remote sensing is the use of satellites to capture information about the surface and atmosphere of the Earth. Remote sensing data are normally stored in raster representations. The global positioning system is a space-based satellite navigation system that provides direct measurement of position and time on the surface of the Earth. Remote sensing images and global positioning system signals can be regarded as primary data sources for the geographic information system.

Spatial Data Service

Various proprietary and public formats for raster and vector representations have been introduced since computers were used for spatial data collection, analysis, and presentation. Plenty of remote sensing images, digital maps, and sensor data form a massive spatial data legacy. On the one hand, they greatly facilitate the progress of using spatial data to tackle scientific and social issues. On the other hand, the heterogeneities caused by the numerous data formats, conceptual models, and software platforms bring huge challenges for data integration and reuse from multiple sources. The Open Geospatial Consortium (OGC) (2016) was formed in 1994 to promote a worldwide

consensus process for developing publicly available interface standards for spatial data. By early 2015, the consortium consists of more than 500 members from industry, government agencies, and academia. Standards developed by OGC have been implemented for promoting interoperability in spatial data collection, sharing, service, and processing. Well-known standards include the Geography Markup Language, Keyhole Markup Language, Web Map Service, Web Feature Service, Web Processing Service, Catalog Service for the Web, Observations and Measurements, etc.

Community efforts such as the OGC service standards offer a solution to publish multisource heterogeneous spatial data legacy on the Web. A number of best practices have emerged in recent years. The OneGeology is an international initiative among the geological surveys across the world. It was launched in 2007, and by early 2015, it has 119 participating member nations. Most members in OneGeology share national and/or regional geological maps through the OGC service standards, such as Web Map Service and Web Feature Service. The OneGeology Portal provides a central node for the various distributed data services. The Portal is open and easy to use. Anyone with an internet browser can view the maps registered on the portal. People can also use the maps in their own applications as many software programs now provide interfaces to access the spatial data services. Another more comprehensive project is the GEO Portal of the Global Earth Observation System of Systems, which is coordinated by the Group on Earth Observations. It acts as a central portal and clearinghouse providing access to spatial data in support of the whole system. The portal provides registry for both data services and standards used in data services. It allows users to discover, browse, edit, create, and save spatial data from members of the Group on Earth Observations across the world.

Another popular spatial data service is the virtual globe, which provides three-dimensional representation of the Earth or another world. It allows users to navigate in a virtual environment by changing the position, viewing angle, and scale.

A virtual globe has the capability to represent various different views on the surface of the Earth by adding spatial data as layers on the surface of a three-dimensional globe. Well-known virtual globes include Google Earth, NASA World Wind, ESRI ArcGlobe, etc. Besides spatial data browsing, most virtual globe programs also enable the functionality of interactions with users. For example, the Google Earth can be extended with many add-ons encoded in the Keyhole Markup Language, such as geological map layers exported from OneGeology.

Open-Source Approaches

There are already widely used free and open-source software programs serving different purposes in spatial handling (Steiniger and Hunter 2013). Those programs can be grouped into a number of categories:

- (1) Standalone desktop geographic information systems such as GRASS GIS, QGIS, and ILWIS
- (2) Mobile and light geographic information systems such as gvSIG Mobile, QGIS for Android, and tangoGPS
- (3) Libraries with capabilities for spatial data processing, such as GeoScript, CGAL, and GDAL
- (4) Data analysis and visualization tools such as GeoVISTA Studio and R and PySAL;
- (5) Spatial database management systems such as PostgreSQL, Ingres Geospatial, and JASPA
- (6) Web-based spatial data publication and processing servers such as GeoServer, MapServer, and 52n WPS
- (7) Web-based spatial data service development frameworks such as OpenLayers, GeoTools, and Leaflet

An international organization, the Open Source Geospatial Foundation, was formed in 2006 to support the collaborative development of open-source geospatial software programs and promote their widespread use.

Companies such as Google, Microsoft, and Yahoo! already provide free map services. One can browse maps on the service website, but the spatial data behind the service is not open. In contrast, the free and open-source spatial data approach requires not only freely available datasets but also details about the data, such as format, conceptual structure, vocabularies used, etc. A well-known open-source spatial data project is the OpenStreetMap, which aims at creating a free editable map of the world. The project was launched in 2004. It adopts a crowdsourcing approach, that is, to solicit contributions from a large community of people. By the middle of 2014, the OpenStreetMap project has more than 1.6 million contributors. Comparing with the maps, the data generated by the OpenStreetMap are considered as the primary output. Due to the crowdsourcing approach, the current data qualities vary across different regions. Besides the OpenStreetMap, there are numerous similar open-source and collaborative spatial data projects addressing the needs of different communities, such as the GeoNames for geographical names and features, the OpenSeaMap for a worldwide nautical chart, and the eBird project for real-time data about bird distribution and abundance.

Open-source spatial data formats have also received increasing attention in recent years, especially Web-based formats. A typical example is GeoJSON, which enables the encoding of simple geospatial features and their attributes using JavaScript Object Notation (JSON). GeoJSON is now supported by various spatial data software packages and libraries, such as OpenLayers, GeoServer, and MapServer. Map services of Google, Yahoo!, and Microsoft also support GeoJSON in their application programming interfaces.

Spatial Intelligence

The Semantic Web brings innovative ideas to the geospatial community. The Semantic Web is a web of data compared to the traditional web of documents. A solid enablement of the Semantic Web is the Linked Data, which is a group of

methodologies and technologies to publish structured data on the Web so they can be annotated, interlinked, and queried to generate useful information. The Web-based capabilities of linking and querying are specific features of the Linked Data, which help people to find patterns from data and use them in scientific or business activities. To make full use of the Linked Data, the geospatial community is developing standards and technologies to (1) transform spatial data into Semantic Web compatible formats such as the Resource Description Framework (RDF), (2) organize and publish the transformed data using triple stores, and (3) explore patterns in the data using new query languages such as GeoSPARQL.

The RDF uses a simple triple structure of subject, predicate, and object. The structure is robust enough to support the linked spatial data consisting of billions of triples. Building on the basis of the RDF, there are a number of specific schemas for representing locations and spatial relationships in triples, such as the GeoSPARQL. Triple stores offer functionalities to manage spatial data RDF triples and query them, which are very similar to what the traditional relational databases are capable. As mentioned above, spatial data have two major sources: conventional data legacy and crowdsourcing data. While technologies are being mature for transforming both of them into triples, the crowdsourcing data provide a more flexible mechanism for the Linked Data approach and data exploration as they are fully open. For example, there are already works done to transform data of the OpenStreetMap and GeoNames into RDF triples. For the pattern exploration, there are already initial results, such as those in the GeoKnow project (Athanasiou et al. 2014). The project built a prototype called GeoKnow Generator which provides functions to link, enrich, query, and visualize RDF triples of spatial data and build lightweight applications addressing specific requests in the actual world.

The linked spatial data is still far from mature yet. More efforts are needed on the annotation and accreditation of shared spatial RDF data, the integration and fusion of them, the efficient RDF query in a big data environment, and innovative ways to visualize and present the results.

Cross-References

- [Geography](#)
- [Socio-spatial Analytics](#)
- [Spatiotemporal Analytics](#)

References

- Athanasiou, S., Hladky, D., Giannopoulos, G., Rojas, A. G., Lehmann, J. (2014). GeoKnow: Making the web an exploratory place for geospatial knowledge. *ERCIM News*, 96. <http://ercim-news.ercim.eu/en96/special/geoknow-making-the-web-an-exploratory-place-for-geo-spatial-knowledge>. Accessed 29 Apr 2016.
- Huisman, O., & de By, R. A. (Eds.). (2009). *Principles of geographic information systems*. Enschede: ITC Educational Textbook Series.
- Open Geospatial Consortium (2016). About OGC. <http://www.opengeospatial.org/ogc>. Accessed 29 Apr 2016.
- Steiniger, S., & Hunter, A. J. S. (2013). The 2012 free and open source GIS software map: A guide to facilitate research, development, and adoption. *Computers, Environment and Urban Systems*, 39, 136–150.

Spatial Econometrics

Giuseppe Arbia

Universita' Cattolica Del Sacro Cuore, Catholic University of the Sacred Heart, Rome, Italy

Spatial Econometrics and Big Data

Spatial econometrics is the branch of scientific knowledge, at the intersection between statistics, economics, and geography, which studies empirically the geographical aspects of economic relationships. The term was coined by the father of the discipline, Jean H. Paelinck, in the general address he delivered to the Annual Meeting of the Dutch Statistical Association in May 1974. The interest in the discipline has recorded a particularly sharp increase in the last two decades which recorded an explosion in the number of applied disciplines interested in the subject and of the number of papers appeared in scientific journals. The major application fields are subjects like regional economics, criminology, public finance, industrial

organization, political sciences, psychology, agricultural economics, health economics, demography, epidemiology, managerial economics, urban planning, education, land use, social sciences, economic development, innovation diffusion, environmental studies, history, labor, resources and energy economics, transportation, food security, real estate, and marketing. But the list of applied disciplines that can benefit from the advances in spatial econometrics is, in fact, a lot longer and likely to further increase in the future. The number of textbooks available to introduce new scholars to the discipline has also raised a lot recently. To the long-standing traditional textbook by Luc Anselin (1988), a list of new volumes were added in the last decade or so (e.g., Arbia 2006, 2014; LeSage and Pace 2009) that can introduce the topic to scholars at various levels of formalization.

The broad field of spatial econometrics can be distinguished into two branches according to the typology of data considered in the empirical analyses. Conventional spatial econometrics treats mainly data aggregated at the level of a real geographical partition, such as countries, regions, counties, or census tracts. This first branch is referred to as *the spatial econometrics of regional data* and represents, to date, the mainstream in the scientific research. The second branch introduces space and spatial relationships in the empirical analysis of individual granular data referring to the single economic agent, thus overcoming the problems connected with data aggregation (see “► [Data Aggregation](#)”). This second branch is termed *spatial microeconometrics* and is emerging in recent years as an important new field of research (Arbia 2016).

Both branches have been interested in the last decades by the big data revolution in terms of the volume and the velocity with which data are becoming more and more available to the scientific community. Data geographically aggregated are more and more available at a very high level of resolution. For instance, the Italian National Statistical Institute releases census information related to about 402,000 census tracts. Many demographic variables are collected by Eurostat at the level of the European regular square (1 km

by-1 km size) lattice grid involving many millions of observations. On the other hand, the availability of very large geo-referenced individual micro-data has also increased dramatically in all fields of economic analysis, making it possible to develop a spatial microeconomic approach which was unconceivable only until few decades ago. For instance, the US Census Bureau provides annual observations for every private sector establishment with payroll and includes approximately 4 million establishments and 70 million employees each year. Examples of this kind can be increasingly found in all branches of economics.

Founding on the mathematical theory of random fields, the basic linear, isotropic (i.e., *directionally invariant*), homoschedastic spatial econometric models are based on the SARAR (acronym for Spatial AutoRegressive with additional AutoRegressive error structure) paradigm. The general formulation of this model is based on the following set of equations:

$$y = \lambda Wy + X\beta_{(1)} + WX\beta_{(2)} + u \quad |\lambda| < 1 \quad (1)$$

$$u = \rho Wu + \varepsilon \quad |\rho| < 1 \quad (2)$$

where y is a vector of n observations of the independent variable, X is an n -by- n matrix of non-stochastic regressors, $\varepsilon|X \approx i.i.d.N(0, \sigma_\varepsilon^2 I_n)$ (with I_n the unitary matrix of dimension n) are the disturbance terms, $\beta_{(1)}$, $\beta_{(2)}$ are vectors of parameters, and λ and ρ scalar parameters to be estimated. The definition of the n -by- n W matrix deserves further explanations. In general the matrix W represents a set of, exogenously given, weights, which depend on the geography of the phenomenon. If data are aggregated at a regional level, the generic entry of the matrix, say $w_{ij} \in \mathcal{W}$, is usually defined by $w_{ij} = \begin{cases} 1 & \text{if } j \in N(i) \\ 0 & \text{otherwise} \end{cases}$ ($N(i)$ being the set of neighbors of location j), with $w_{ii} = 0$ by definition. Conversely, if data represent granular observations on the single economic agent, the W matrix is based on the information about the (physical or economic) pairwise

interpoint distances. In this case, many different definitions are possible considering, e.g., (i) an inverse function of the interpoint distances d_{ij} , e.g., $w_{ij} = d_{ij}^{-\alpha}$; $\alpha > 0$; (ii) a threshold criterion expressed in a binary form by $w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < d^* \\ 0 & \text{otherwise} \end{cases}$, d^* being the threshold distance; (iii) a combination of threshold and inverse distance definition such that $w_{ij} = \begin{cases} d_{ij}^{-\alpha} & \text{if } d_{ij} < d^* \\ 0 & \text{otherwise} \end{cases}$; and (iv) a nearest neighbors definition. Equation 1 considers the spatially lagged variable of the dependent variable y (the term WY) as one of the regressors and may also contain spatially lagged variables of some or all of the exogenous variables (the term WX). Equation 2 considers a spatial autoregressive model for the stochastic disturbances (the term Wu).

The SARAR model represents the benchmark for the analysis of both regional and individual microgeographical data. There are, however, some important differences in the two cases. Indeed, when dealing with regional data, almost invariably, the spatial units constitute a complete cross section of territorial units with no missing data, variables are observed directly, there is no uncertainty on the spatial observations that are free from measurement error, and the location of the regions is perfectly known. In contrast, granular spatial microdata quite often present different forms of imperfections: they are often based on a sample drawn from a population of spatial locations, some data are missing, some variable only proxy the target variables, and they almost invariably contain both attribute and locational errors (see “► [Big Data Quality](#)”).

Many possible alternatives have been proposed to estimate the parameters of model (1–2) (see Arbia 2014 for a review). A maximum likelihood (ML) approach assuming normality of the residuals guarantees the optimal properties of the estimators, but, since no closed-form solution is generally available, the solutions have to be obtained numerically raising severe problems of computing time, storage, and accuracy. Alternatively, the generalized method of moments

(GMM) procedures that have been proposed do not require any distributional assumptions and may reduce (although not completely eliminate) the computational problems in the presence of very large databases and very dense W matrices. These estimators, however, are not fully efficient. Further models have been suggested in the literature to overcome the limits of the basic SARAR model, considering methodological alternatives to remove the (often unrealistic) hypotheses of isotropy, linearity, and homoschedasticity on which they are based (Arbia 2014 for a review). Methods and models are also available for the analysis of spatiotemporal econometric data (Baltagi 2013) (see “► [Spatiotemporal Analytics](#)”).

The estimation of both the regional and the microeconomic models may encounter severe computational problems connected with the dimension of the dataset. Indeed, both the ML and the GMM estimation procedures require repeated inversions of an n -by- n matrix expressed as some function of the W matrix. If n is very large, this operation could become highly demanding if not prohibitive. A way out, employed for years in the literature, consisted of exploiting an eigenvalue decomposition of the matrices involved, a solution which, however, does not completely eliminate the accuracy problems if n is very large because the spectral decomposition in very large matrices is the outcome of an approximation. Many studies report that the computation of eigenvalues by standard subroutines for general nonsymmetric matrices may be highly inaccurate already for relatively small sample sizes ($n > 400$). The accuracy improves if the matrix is symmetric, which, unfortunately, is not always the case with spatial econometrics models. Many other approximations were proposed, but none entirely satisfactory especially when the W matrices are very dense. The computational issues connected with the estimation of spatial econometric models are doomed to become more and more severe in the future even with the increasing power of computer machines and the diffusion of parallel processing (see “► [Parallel Processing](#)”) because

the availability of very large databases is also increasing at an accelerated speed. Apart from a large number of attempts to simplify the problem computationally, some of the most recent literature has concentrated on the specification of alternative models that are computationally simpler. In this respect the three most relevant methods that can be used for big data in spatial econometrics are the *matrix exponential spatial specification (MESS)*, the *unilateral approximation*, and the *bivariate marginal likelihood approach* (see Arbia 2014 for a review).

Conclusions

Spatial econometrics is a rapidly changing discipline also due to the formidable explosion of data availability and their diffusion in all spheres of human society. Under this respect the use of satellite data and the introduction of new sophisticated positioning devices together with the widespread access to individual granular data deriving from archives, from social networks, crowdsourcing, and other sources have the potential to revolutionize the way in which econometric modelling of spatial data will be approached in the future. Under this respect, in the future we will progressively observe a transition from economic phenomena that are modelled on a discrete to phenomena which are observed on a continuous in space and time and that will need a completely novel set of tools. This is certainly true for many phenomena that are intrinsically continuous in space and that were observed so far on a discrete only due to our limitations in the observational tools (e.g., environmental variables), but also for phenomena characterized by spatial discontinuities, like those observed in transportation or health studies, just to make few examples. Under this point of view, spatial econometrics will benefit in the future from the cross-contamination with techniques developed for the analysis of continuous spatial data and with other useful tools that could be borrowed from physics.

Cross-References

- [Big Data Quality](#)
- [Data Aggregation](#)
- [Parallel Processing](#)
- [Socio-spatial Analytics](#)
- [Spatiotemporal Analytics](#)

Further Reading

- Anselin, L. (1988). *Spatial econometrics, methods and models*. Dordrecht: Kluwer Academic.
- Arbia, G. (2006). *Spatial econometrics: Statistical foundations and applications to regional convergence*. Heidelberg: Springer.
- Arbia, G. (2014). *A primer for spatial econometrics*. Basingstoke: Palgrave-MacMillan.
- Arbia, G. (2016). Spatial econometrics. *Foundations and Trends in Econometrics*, 8(3–4), 145–265.
- Baltagi, B. (2013). *Econometric analysis of panel data* (5th ed.). New York: Wiley.
- LeSage, J., & Pace, K. (2009). *Introduction to spatial econometrics*. Boca Raton: Chapman and Hall/CRC Press.

Spatial Scientometrics

Song Gao
 Department of Geography, University of
 California, Santa Barbara, CA, USA
 Department of Geography, University of
 Wisconsin-Madison, Madison, WI, USA

Synonyms

[Geospatial scientometrics](#)

Definition/Introduction

The research field of scientometrics (or bibliometrics) is concerned with measuring and analyzing science, with the aim of quantifying a publication, a journal, or a discipline's structure, impact, change, and interrelations. The spatial dimension (e.g., location, place, proximity) of

science has been added into account since research activities usually start from a certain region or several places in the world and then spread to other places, thus displaying spatiotemporal patterns. The analysis of spatial aspects of the science system is composed of spatial scientometrics (Frenken et al. 2009), which address the studies of geospatial distribution patterns on scientific activities, domain interactions, co-publications, citations, academic mobility, and so forth. The increasing availability of large-scale research metadata repositories in the big data age and the advancement in geospatial information technologies have enabled geospatial big data analytics for the quantitative study of science.

Main Research Topics

The earliest spatial scientometrics studies date back to 1970s. Researchers analyzed the distribution of worldwide science productivity by region and country. Later on, the availability of more detailed affiliation address information, and geographic coordinate data offers the possibility to investigate the role of physical distance in collaborative knowledge production. And the “spatial” dimension can refer to not only the “geographic space” but also the “cyberspace.” The book *Atlas of Science: Visualizing What We Know* collected a series of visual maps in cyberspace for navigating the dynamic structure of science and technology (Börner 2010). According to the research framework for spatial scientometrics proposed by Frenken et al. (2009), there are at least three main topics addressed in this research domain: (1) *spatial distribution*, it studies the location arrangement of different scientific activities including research collaborations, publications, and citations across the Earth's surface. Whether geographic concentration or clustered patterns can bring advantages in scientific knowledge production is an important research issue in spatial scientometrics. (2) *Spatial bias*, it refers to those uneven spatial distributions on the scientific activities and their structure because of the limits on research funding, intellectual property, equipment, language, and so on. One prominent spatial

bias is that researchers collaborate domestically more frequently than internationally, which might also be influenced by the number of researchers in a country. Another spatial bias is that collaborative articles from nearby research organizations are more likely to be cited than articles from research organizations further away within the same country. But there is a positive effect of international co-publications on citation impact compared with domestic co-publications. Such patterns might change with the increasing accessibility of crowdsourced or open-sourced bibliographic databases. Regarding researchers' trajectory or academic mobility patterns, they are also highly skew distributed across countries. Recent interests arise in the analysis of the origin patterns of regional or international conference participants. (3) *Citation impact*, it attracts much attention in the scientometrics studies. In academia, the number of citations is an important criterion to estimate the impact of a scientific publication, a journal, or a scientist. Spatial scientometrics researchers study and measure the geospatial distributions and impacts of citations for scientific publications and knowledge production.

Key Techniques and Analysis Methods

In order to analyze the geospatial distribution and interaction patterns of scientific activities in scientometrics studies, one important task is to get the location information of publications or research activities. There are two types of location information: (1) *place names* at different geopolitical scales (e.g., city, state, country, region) and (2) *geographic coordinates* (i.e., latitude and longitude). The place information can usually be retrieved from the affiliation information in standard bibliographic databases such as *Thomson Reuters Web of Science* or *Elsevier Scopus*. But the geographic coordinate information is not directly available in those databases. Additional processing techniques “georeferencing” which assigns a geographic coordinate to a place-name and “geocoding” which converts an address text into a geographic coordinate are required to generate the coordinate information for a publication,

a citation, or a researcher. Popular geocoding tools include *Google Maps Geocoding API* and *ArcGIS Online Geocoding Service*.

After getting the coordinate information, a variety of statistical analysis and mapping/geovisualization techniques can be employed for spatial scientometrics analyses (Gao et al. 2013). A simplistic approach showing the spatial distribution pattern is to map the affiliation location of publications or citations or to aggregate the affiliation locations to the administrative places (e.g., city or country boundaries). Another method is to use the kernel density estimation (KDE) mapping to identify the “hotspot regions” in the geography of science (Bornmann and Waltman 2011). The KDE mapping has been widely used in spatial analysis to characterize a smooth density surface that shows the geographic clustering of point or line features. The two-dimensional KDE can identify the regions of citation clusters for each cited paper by considering both the quantity of citations and the area of geographical space, compared to the single-point representation which may neglect the multiple citations in the same location. Moreover, the concept of geographic proximity (distance) is widely used to quantify the spatial patterns of co-publications and citations. In addition, the socioeconomic factors that affect the scientific interactions have also been addressed. Boschma (2005) proposed a proximity framework of physical, cognitive, social, and institutional forms to study the scientific interaction patterns. Researchers studied the relationship between each proximity and citation impact by controlling other proximity variables. Also, the change of author affiliations over time adds complexity to the network analysis of universities. The approach with thematic, spatial, and similarity operators has been studied in the GIScience community to address this challenging issue.

When measuring the citation impact of a publication, a journal, or a scientist, traditional approaches purely counting the number of citations do not take into account the geospatial and temporal impact of the evaluating target. The spatial distribution of citations could be different even for publications with the same number of citations. Similarly, some work may be relevant

and cited for decades, while other contributions only have a short-term impact. Therefore, Gao et al. (2013) proposed a theoretical and novel analytical spatial scientometrics framework which employs spatiotemporal KDE, cartograms, distance distribution curves, and spatial point patterns to evaluate the spatiotemporal citation impacts for scientific publications and researchers. Three geospatial citation impact indices ($S_{\text{institution}}$ index, S_{city} index, and S_{country} index) were developed to evaluate an individual scientist's geospatial citation impact, which complement traditional nonspatial measures such as h-index and g-index.

Challenges in the Big Data Age

Considering the three V's characteristics (volume, velocity, and variety) of big data, there are many challenges in big-data-driven (spatial) scientometrics studies. These challenges require both computationally intensive processing and careful research design (Bratt et al. 2017). First, the author names, affiliation trajectory, and institution names and locations often need to be disambiguated and uniquely identified. Second, the heterogeneous formats (i.e., structured, semi-structured, unstructured) of bibliographic data might be incredibly varied and cannot fit into a single spreadsheet or a database application. Moreover, the metadata standards are inconsistent across multiple sources and may change over time. All the abovementioned challenges can affect the validity and reliability of (spatial) scientometrics studies. The uncertainty or sensitivity analyses need to be included in the data processing and analytical workflows.

Conclusion

Spatial scientometrics involves the studies of spatial patterns, impacts, and trends of scientific activities (e.g., co-publication, citation, academic mobility). In the new era, because of the increasing availability of digital bibliographic databases and open data initiatives, researchers from multiple domains can contribute various qualitative,

quantitative, and computational approaches and technologies into the spatial scientometrics analyses. The spatial scientometrics is still an infant interdisciplinary field with the support of spatial analysis, information science and statistic methodologies. New data sources and measurements to evaluate the excellence in the geography of science are emerging in the age of big data.

Further Reading

- Börner, K. (2010). *Atlas of science: Visualizing what we know*. Cambridge: The MIT Press.
- Bornmann, L., & Waltman, L. (2011). The detection of "hot regions" in the geography of science – A visualization approach by using density maps. *Journal of Informetrics*, 5(4), 547–553.
- Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies*, 39(1), 61–74.
- Bratt, S., Hemsley, J., Qin, J., & Costa, M. (2017). Big data, big metadata and quantitative study of science: A workflow model for big scientometrics. *Proceedings of the Association for Information Science and Technology*, 54(1), 36–45.
- Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3(3), 222–232.
- Gao, S., Hu, Y., Janowicz, K., & McKenzie, G. (2013, November). A spatiotemporal scientometrics framework for exploring the citation impact of publications and scientists. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 204–213). Orlando, Florida, USA: ACM.

Spatiotemporal Analytics

Tao Cheng and James Haworth
SpaceTimeLab, University College London,
London, UK

Spatiotemporal analytics or space-time analytics (STA) is the use of integrated space-time thinking and computation to discover insights from geolocated and time-stamped data. This involves extracting unknown and implicit relationships, structures, trends, or patterns from massive datasets collected at multiple locations and times

that make up space-time (ST) series. Examples of such datasets include daily temperature series at meteorological stations, street-level crime counts in world capital cities, and daily traffic flows on urban roads. The toolkit of STA includes exploratory ST data analysis (ESTDA) and visualization, spatiotemporal modeling, prediction, classification and clustering (profiling), and simulation, which are developed based upon the latest progress in spatial and temporal analysis.

STA starts with ESTDA, which is used to explore patterns and relationships in ST data. This ranges from ST data visualization and mapping to geovisual analytics and statistical hypothesis testing. ST data visualization explores the patterns hidden in the large ST datasets using visualization, animation, and interactive techniques. This includes conventional 2D maps and graphs alongside advanced 3D visualizations. The 3D space-time cube, proposed by Hägerstrand (1970), is an important tool in STA. It consists of two dimensions of geographic locations on a horizontal plane and a time dimension in the vertical plane (or axis). The space-time cube is used to visualize trajectories of objects in 3D space-time dimension, or “space-time paths,” but can also show hotspots, isosurfaces, and densities (Cheng et al. 2013; Demsar et al. 2015).

In STA, ST data visualization is undertaken as part of an iterative process involving information gathering, data preprocessing, knowledge representation, and decision-making, which is known as geovisual analytics (Andrienko et al. 2007). Geovisual analytics is an extension of visual analytics, which is becoming more important to many disciplines including scientific research, business enterprise, and other areas that face problems of an overwhelming avalanche of data. First, ST data are visualized to reveal basic patterns, and then users will use their perception (intuition) to gain insights from the images produced. Insights generated are then transformed into knowledge. This knowledge can be used to generate hypotheses and carry out further ESTDA, the results of which will be visualized for presentation, and further knowledge generation. Geovisual analytics is an integrated approach to combining ST data

visualization with expert knowledge and data analysis.

Hand in hand with geovisual analytics go the statistical ESTDA tools of STA. Particularly central to STA is the concept of spatiotemporal dependence. To paraphrase Tobler’s first law of geography (Tobler 1970), an observation from nature is that near things tend to be more similar than distant things both in space and in time. A space-time series may exhibit ST dependence, which describes its evolution over space and time. If the ST dependence in a dataset can be modeled, then one can make predictions of future values of the series. ST dependence can be quantitatively measured using ST autocorrelation indices such as the ST autocorrelation function (Cheng et al. 2011) and the ST (semi)variogram (Griffith and Heuvelink 2009), which are key tools of ESTDA. Also important are tools for measuring ST heterogeneity, whereby global patterns of ST autocorrelation are violated at the local level. When dealing with point patterns, tests for ST clustering or ST density estimation may be used.

ESTDA helps to reveal the most appropriate method for STA, which varies depending on the data type and objective. Alongside visualization, the core tasks of STA are predictive modeling, clustering/profiling, and simulation. Predictive modeling involves using past values of a ST series (and possible covariates) to forecast future values. Depending on the data, predictive modeling may involve either classification, whereby the desired output is two or more classes, or regression, whereby the desired output is continuous. These tasks are referred to as supervised learning as the desired output is known. Predictive modeling methods can be separated into two broad categories: statistical and machine learning approaches. Statistical methods are generally adaptations of existing models from the fields of time series analysis, spatial analysis, and econometrics to deal with spatiotemporal data. Some of the methods commonly used in the literature include space-time autoregressive integrated moving average (STARIMA) models (Pfeifer and Deutsch 1980) and variants, multiple ARIMA models, space-time geostatistical models (Heuvelink and

Griffith 2010), spatial panel data models (Elhorst 2003), geographically and temporally weighted regression (Huang et al. 2010; Fotheringham et al. 2015), and eigenvector spatial filtering (Patuelli et al. 2009). More recently, Bayesian hierarchical models have become popular due to their ability to capture spatial, temporal, and spatiotemporal effects (Blangiardo and Cameletti 2015).

The aforementioned methods tend to rely on strong statistical assumptions and can be difficult to fit to large datasets. Increasingly, researchers and practitioners are turning toward machine learning and data mining methods that are better equipped to deal with the heterogeneous, nonlinear, and multiscale properties of big ST data. Artificial neural networks (ANNs), support vector machines (SVMs), and random forests (RFs) are now being successfully applied to ST predictive modeling problems (Kanevski et al. 2009). ANNs are a family of nonparametric methods for function approximation that have been shown to be very powerful tools in many application domains. They are inspired by the observation that biological learning is governed by a complex set of interconnected neurons. Although individual neurons may be simple in structure, their interconnections allow them to perform complex tasks such as pattern recognition and classification. SVMs are a set of supervised learning methods originally devised for classification tasks that are based on the principles of statistical learning theory (Vapnik 1999). SVMs use a linear algorithm to find a solution to classification or regression problems that is linear in a feature space and nonlinear in the input space. This is accomplished using a kernel function. SVMs have many advantages: (1) they have a globally optimal solution; (2) they have a built-in capacity to deal with noisy data; and (3) they can model high-dimensional data efficiently. These advantages have made SVMs, along with other kernel methods, a very important tool in STA. RFs are ensembles of decision trees. They work on the premise that the mode (classification) or average (regression) of a large number of trees trained on the same data will tend towards an optimal solution. RFs have achieved performance comparable to SVMs and ANNs and are becoming common in STA.

The aforementioned predictive modeling methods assume knowledge of the desired output. Often we will know little about an ST dataset and may wish to uncover hidden structure in the data. This is known as unsupervised learning and is addressed using clustering methods. Clustering involves grouping unlabeled objects that share similar characteristics. The goal is to maximize the intraclass similarity and minimize the interclass similarity. Widely used spatial clustering techniques, e.g., K-means and K-medoids, have been extended to spatiotemporal clustering problems. Initial research on spatial clustering has focused on point data with popular algorithms such as DBSCAN and BIRCH. However, designing an effective ST clustering algorithm is a difficult task because it must account for the dynamics of a phenomenon in space and time.

Very few clustering algorithms consider the spatial, temporal, and thematic attributes seamlessly and simultaneously. Capturing the dynamicity in the data is the most difficult challenge in ST clustering, which is the reason that traditional clustering algorithms, in which the clustering is carried out on a cross section of the phenomenon, cannot be directly applied to ST phenomena. The arbitrarily chosen temporal intervals may not capture the real dynamics of the phenomena since they only consider the thematic values at the same time, which cannot capture the influence of flow (i.e., time lag phenomena). It is only recently that this has been attempted. ST-DBSCAN is one method that has been developed and applied to clustering ST data (Birant and Kut 2007). Spatiotemporal scan statistics (STSS) is a clustering technique that was originally devised to detect disease outbreaks (Neill 2008). The goal is to automatically detect regions of space that are “anomalous,” “unexpected,” or otherwise “interesting.” Spatial and temporal proximities are exploited by scanning the entire study area via overlapping space-time regions (STRs). Each STR represents a possible disease outbreak with a geometrical shape which is either a cylinder or rectangular prism. The base corresponds to the spatial dimension and the height corresponds to the temporal dimension. The dimensions of the STR are allowed to vary in order to detect outbreaks of varying sizes.

The final task of STA discussed here is simulation, which involves the development of models for simulating complex ST processes. Two common methods are cellular automata (CA) and agent-based modeling (ABM) (Batty 2007). In CA, a spatial region is divided into cells that have certain states. The probability of a cell changing from one state to another is affected by the state of surrounding cells at the same or previous times. In ABMs, agents are constructed that have certain behaviors that determine their interaction with their environment and other agents. In both model types, the aim is to study emergent behavior from small-scale interactions. Simulation models have been applied to study many phenomena including traffic congestion, urban change, emergency evacuation and vegetation dynamics, and policing and security. If properly calibrated, simulation models can be used to predict ST processes over long time periods and to develop and test theories. However, the principal issue with such methods is validation against real data, which is only recently being addressed (Wise and Cheng 2016).

Using the toolkit of STA, the researcher can uncover insights into their ST data that they may otherwise miss and make their data work for them, thus realizing its potential whether it be for business or scientific research.

Further Reading

- Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M. J., MacEachren, A., & Wrobel, S. (2007). Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, 21, 839–857.
- Batty, M. (2007). *Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals*. London: The MIT Press.
- Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowledge Engineering Intelligent Data Mining*, 60, 208–221. <https://doi.org/10.1016/j.datak.2006.01.013>.
- Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R – INLA* (1st ed.). Chichester: Wiley.
- Cheng, T., Haworth, J., & Wang, J. (2011). Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*. <https://doi.org/10.1007/s10109-011-0149-5>.
- Cheng, T., Tanaksaranond, G., Brunsdon, C., & Haworth, J. (2013). Exploratory visualisation of congestion evolutions on urban transport networks. *Transportation Research Part C: Emerging Technologies*, 36, 296–306. <https://doi.org/10.1016/j.trc.2013.09.001>.
- Demsar, U., Buchin, K., van Loon, E. E., & Shamoun-Baranes, J. (2015). Stacked space-time densities: A geovisualisation approach to explore dynamics of space use over time. *GeoInformatica*, 19, 85–115. doi:10.1007/s10707-014-0207-5.
- Elhorst, J. P. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review*, 26, 244–268. <https://doi.org/10.1177/0160017603253791>.
- Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Geographical and temporal weighted regression (GTWR). *Geographical Analysis*, 47, 431–452. <https://doi.org/10.1111/gean.12071>.
- Griffith, D. A., & Heuvelink, G. B. (2009, June). *Deriving space-time variograms from space-time autoregressive (STAR) model specifications*. In: StatGIS 2009 Conference, Milos, Greece.
- Hägerstrand, T. (1970). What about people in regional science? *Papers in Regional Science*, 24, 7–24. <https://doi.org/10.1111/j.1435-5597.1970.tb01464.x>.
- Heuvelink, G. B. M., & Griffith, D. A. (2010). Space-time geostatistics for geography: A case study of radiation monitoring across parts of Germany. *Geographical Analysis*, 42, 161–179.
- Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24, 383–401. <https://doi.org/10.1080/13658810802672469>.
- Kanevski, M., Timonin, V., & Pozdnukhov, A. (2009). *Machine learning for spatial environmental data: Theory, applications, and software*. Har/Cdr. (Ed.), EFPL Press.
- Neill, D. B. (2008). Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 25(3), 498–517.
- Patuelli, R., Griffith, D. A., Tiefelsdorf, M., & Nijkamp, P. (2009). *Spatial filtering and eigenvector stability: Space-time models for German unemployment data*. Quad. Della Fac. Sci. Econ. Dell'Università Lugano.
- Pfeifer, P. E., & Deutsch, S. J. (1980). A three-stage iterative procedure for space-time modelling. *Technometrics*, 22, 35–47.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240. <https://doi.org/10.2307/143141>.
- Vapnik, V. (1999). *The nature of statistical learning theory* (2nd ed.). New York: Springer.
- Wise, S. C., & Cheng, T. (2016). How officers create guardianship: An agent-based model of policing. *Transactions in GIS*, 20, 790–806. <https://doi.org/10.1111/tgis.12173>.

Speech Processing

► Voice User Interaction

Speech Recognition

► Voice User Interaction

Standardization

Travis Loux
Department of Epidemiology and Biostatistics,
College for Public Health and Social Justice, Saint
Louis University, St. Louis, MO, USA

Definition/Introduction

In many big data applications, the data was not collected through a formally designed study, but through available means. The data is often observational (with no randomization mechanism) or incomplete (a convenience sample rather than the full population) (Fung 2014). Thus, subgroups within the data and the data set as a whole may not be representative of the appropriate population, leaving analyses based on such data open to biases. Data standardization is the process of scaling a data set to be comparable to a reference

distribution. In practice, standardization is used to equate two or more groups within a sample on a subset of variables (internal standardization) or to equate a sample to an external source, such as another sample or a known population (external standardization). Roughly speaking, internal standardization is used to perform causal inferences between the two groups, while external standardization is used to generalize results from a sample to a population. Done carefully, standardization allows analysts to make inferences and generalizations they would be unable to make with basic statistical comparisons.

A Simple Example

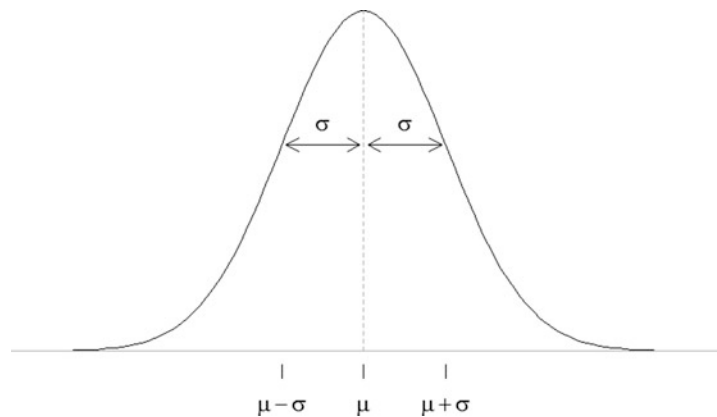
One of the most commonly used standardization tools, and an instructive starting point, is the *standard normal distribution*, commonly denoted as the *Z* distribution. If a variable *Y* follows a normal distribution with mean μ and standard deviation σ (Fig. 1), it can be standardized through the formula

$$Z = \frac{Y - \mu}{\sigma}$$

The resulting *Z* variable will also be normally distributed, but will have mean 0 and standard deviation 1. This is a standardization process because *Z* will have the same distribution regardless of the initial values of μ and σ , meaning any

Standardization,

Fig. 1 The density function of a normal distribution with mean μ and standard deviation σ



two normal distributions can be transformed to the same scale.

The initial motivation for standardizing normal distributions was a computational one: it is difficult to compute an area under a normal curve using pen-and-paper methods. The ability to rescale any normal distribution to a standard one meant that a single table of numbers could provide all the information necessary to compute areas for any normal distribution.

For example, IQ scores are approximately normal with a mean of 100 and a standard deviation of 15. To find the probability of having an IQ score above 120, the IQ distribution can be standardized, reducing the problem to one involving the standard normal distribution:

$$\begin{aligned} P(IQ > 120) &= P\left(\frac{IQ - 100}{15} > \frac{120 - 100}{15}\right) \\ &= P(Z > 1.33) \end{aligned}$$

Similarly, the heights of adult males in the USA are approximately normally distributed with a mean of 69 in. and a standard deviation of 3 in. To find the probability of being taller than 73 in., one can follow a similar procedure:

$$\begin{aligned} P(HT > 73) &= P\left(\frac{HT - 69}{3} > \frac{73 - 69}{3}\right) \\ &= P(Z > 1.33) \end{aligned}$$

Both solutions require only the Z distribution; these problems in very different contexts can be solved by referring to the same standardized scale.

Beyond Z

Standardization can be performed on data that is not normally distributed. In this case, the resulting standardized score $(y - \mu)/\sigma$ can generally be interpreted as the number of standard deviations from the mean above which the observation y lies, with positive standardized scores meaning y is greater than μ and negative scores meaning y is less than μ .

Another basic standardization process is *normalization* (though this term has multiple meanings in data-centric fields). Normalization rescales a data set so that all values are between 0 and 1, using the formula

$$y' = \frac{y - y_{\min}}{y_{\max} - y_{\min}}$$

where $y_{\min}$ and $y_{\max}$ are the minimum and maximum values of the data set, respectively. The resulting normalized values y' range from 0 to 1. Once a data set has been normalized, values can be compared within data sets based on relative standing or across data sets based on location relative to the range of the respective data sets.

Direct Versus Indirect Standardization

Standardization has a long history in the study of epidemiology (e.g., Miettinen 1972). Within the field, there is a distinction made between direct standardization and indirect standardization. *Direct standardization* scales an outcome from a study sample to estimate rates in a reference population, while *indirect standardization* scales an outcome from a reference population to estimate rates in the study sample. In both cases, the study sample and reference population are stratified in similar ways (e.g., using the same age strata cut points).

In direct standardization, the risk or rate of outcome within each stratum of the study sample is calculated, then multiplied by the stratum size within the reference population to estimate the number of events expected to occur in the relevant stratum of the reference population. These expected stratum-specific counts are then totaled over all strata and divided by the total reference population size. Using the notation from Table 1 below, the direct standardized rate in the reference

population is $\left(\sum_i m_i \frac{x_i}{n_i}\right)/M$, resulting in an estimated rate in the reference population.

Indirect standardization reverses this process, by applying stratum-specific outcome rates from

Standardization, Table 1 Example table for direct and indirect standardization

Stratum	Study population		Reference population	
	Events	Size	Events	Size
1	x_1	n_1	y_1	m_1
2	x_2	n_2	y_2	m_2
...	...	...	...	...
k	x_k	n_k	y_k	m_k
Total		N		M

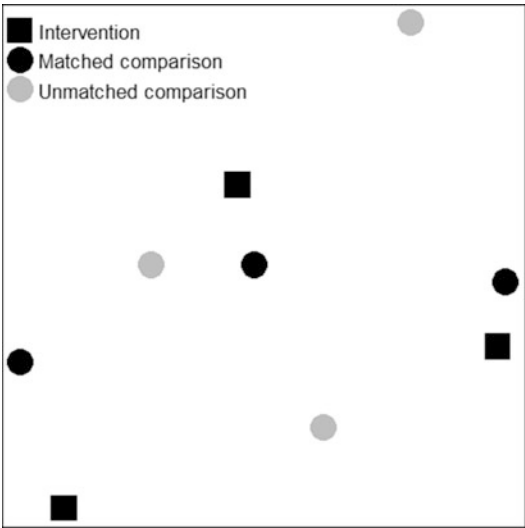
the reference population to the study sample strata and taking a stratum-size weighted average across the strata of the study sample. In Table 1, the indirect standardized rate in the study sample is $\left(\sum_i n_i \frac{y_i}{m_i}\right)/N$. The result of indirect standardization is the outcome rate that would have been expected in the study sample if the individuals experienced the outcome at the same rate as the reference population.

With larger data sets, simple stratification like the methods discussed above can be improved upon by using more variables, yielding more fully standardized results, and more observations, allowing for finer strata.

Internal Standardization

Internal standardization attempts to balance two or more subsets of a data set on a set of baseline variables and is commonly used in causal inference analyses. Common approaches to internal standardization include finding similar observations across the subsets through matching or weighting observations within the subsets so the baseline variables mirror a reference distribution.

There are numerous matching algorithms (Stuart 2010), though most follow a similar framework. In “greedy” or “nearest neighbor” one-to-one (1:1) matching, an observation from one group (usually the intervention or exposure group in a causal inference setting) is randomly selected, and a similar observation from the comparison group is found and paired to the intervention observation. This process is repeated until all observations in the intervention group have



Standardization, Fig. 2 Matching of a hypothetical data set based on closeness of the variables represented on the horizontal and vertical axes

been paired to an observation on the comparison group. The result is a subset of the comparison group in which each individual looks similar to an observation in the intervention group, effectively standardizing the comparison group to the distribution of the intervention group (Fig. 2). Variations on this concept include matching two or more comparison observations to each intervention observation (1:2 or 1:k matching).

In contrast to greedy matching, optimal matching attempts to find the best comparison subset by using a global measure of imbalance between intervention and comparison groups and has been found to yield significant improvements over greedy matching (Rosenbaum 1989). Optimal matching is far more computationally intensive with regards to both memory/storage and

time, and may be infeasible in some big data settings without substantial resources. Other advances in matching include matching for more than two groups (e.g., Rassen et al. 2013). As an additional benefit of standardizing groups, matching also makes resulting analytical conclusions less susceptible to model misspecifications (Ho et al. 2007).

An alternative, or in some cases complementary, approach to matching is weighting. Standardization weighting begins with modeling the probability of subgroup membership based on a set of relevant variables. Weights are then defined as the ratio of the probabilities of group membership. For example, suppose an analyst wants to standardize group $G = B$ to group $G = A$. Then the weight for observation i in Group B is $P(G = A \mid X = x_i)/P(G = B \mid X = x_i)$, where X contains the variables to be standardized (Sato and Matsuyama 2003). Group membership probabilities can be estimated using any standard classification algorithm such as logistic regression or neural networks.

In the simple intervention/comparison setting, this group membership probability is called a propensity score (Rosenbaum and Rubin 1983) and commonly denoted as $e(x_i) = P(T = 1 \mid X = x_i)$, where T is an indicator for intervention. To standardize the comparison group to the intervention group, observation i in the comparison group gets the weight $e(x_i)/(1 - e(x_i))$. Alternatively, one could standardize both the intervention and comparison group to the full sample distribution. To perform this analysis, observations in the intervention group get weight $1/e(x_i)$ while observations in the comparison group get weight $1/(1 - e(x_i))$. In more complex settings analysts can use the generalized propensity score (Imai and Dyk 2004) to obtain group membership probabilities.

As a concrete example, suppose in a data set there is a subgroup of $n = 1000$ for which $e(x_i) = 0.20$. Within this subgroup, there will be 200 intervention and 800 comparison observations. Weighting each intervention observation by $1/e(x_i) = 1/0.20 = 5$ will yield a weighted sample size of $200 \times 5 = 1000$, while weighting each comparison observation by $1/(1 - e(x_i)) = 1/0.80 = 1.25$ will yield a weighted sample size of

$800 \times 1.25 = 1000$. The weighted distributions of the variables included in X will also match the full sample distribution. Thus, both the intervention and comparison groups are standardized to the full sample.

External Standardization

External standardization is used to scale a sample to a reference data source in order to account for selection bias. Common applications of external standardization include adjusting nonrandom samples to match a well-defined population, for example, the US voting population, and generalizing results from randomized trials.

An increasingly popular tool for standardizing a nonrandom sample to a target population is multilevel regression with poststratification (MRP). To begin MRP, a multilevel regression model is applied to a data set. The predicted values from this model are then weighted in proportion to the distribution of the generating predictors in the target population. Wang, Rothschild, Goel, and Gelman (2014) collected data on voting history and intent from a sample of X box users leading up to the 2012 US presidential election. Fitted values from the regression model were weighted to match US electorate on demographics, political party identification, and 2008 presidential vote. In a retrospective analysis, this approach predicted national voting results within one percentage point. In other uses of MRP, Ghitza and Gelman (2013) and Zhang et al. (2015) used results from large surveys to weight specific small, hard-to-sample populations (electoral subgroups in Ghitza and Gelman (2013) and Missouri counties in Zhang et al. (2015)).

Though analysis for intervention effects requires internal standardization between intervention and comparison groups for causal inferences of average effect, external standardization is necessary to obtain population estimates if the effects vary across individuals (called heterogeneous effects). Cole and Stuart (2010) adapted propensity score weighting to generalize the estimated treatment effects from a clinical trial of HIV treatment to the full US HIV-positive population

as estimated by the Center for Disease Control and Prevention. The two data sets were combined with a selection variable indicating selection into the trial from the general population. Cole and colleagues then use propensity score weighting, replacing the intervention indicator with the selection indicator, to standardize the clinical trial sample to the larger HIV-positive population. Stuart, Cole, Bradshaw, and Leaf (2011) further developed these methods in the context of a school-level behavioral intervention study and used propensity score weighting to evaluate the magnitude of selection bias. Rudolph, Diaz, Rosenblum, and Stuart (2014) investigated the use of other internal standardization techniques for estimating population intervention effects.

Conclusion

Standardization can be used to alleviate some of the pitfalls of working with big data (Fung 2014). Since big data is usually observational in nature, causal inferences cannot be made from basic between-group comparisons. Internal standardization will equate intervention or exposure groups on a set of baseline variables. This procedure ensures comparability between the two groups on these measures and excludes them as potential causal explanations. In addition, big data is often not complete and may have serious selection biases, meaning certain types of observations may be systematically less likely to appear in the data set. A naive analysis of such data will yield results that reflect this disparity and do not accurately represent the broader population. External standardization can be used to poststratify, weight, or otherwise equate a data set with a known population distribution, e.g., from the US Census Bureau. The resulting conclusions may then be more representative of the full population and more easily generalizable.

Cross-References

- Association Versus Causation
- Big Data Quality

- Correlation Versus Causation
- Data Quality Management
- Demographic Data
- Regression

Further Reading

- Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, 172(1), 107–115. <https://doi.org/10.1093/aje/kwq084>.
- Fung, K. (2014). Toward a more useful definition of Big Data. Retrieved from <http://junkcharts.typepad.com/numbersruleyourworld/2014/03/toward-a-more-useful-definition-of-big-data.html>.
- Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3), 762–776. <https://doi.org/10.1111/ajps.12004>.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199–236. <https://doi.org/10.1093/pan/mpl013>.
- Imai, K., & Dyk, D. a. v. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467), 854–866. <https://doi.org/10.1198/016214504000001187>.
- Miettinen, O. S. (1972). Standardization of risk ratios. *American Journal of Epidemiology*, 96(6), 383–388.
- Rassen, J. A., Shelat, A. A., Franklin, J. M., Glynn, R. J., Solomon, D. H., & Schneeweiss, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology*, 24(3), 401–409. <https://doi.org/10.1097/EDE.0b013e318289dedf>.
- Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408), 1024–1032. <https://doi.org/10.1080/01621459.1989.10478868>.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. <https://doi.org/10.1093/biomet/70.1.41>.
- Rudolph, K. E., Diaz, I., Rosenblum, M., & Stuart, E. A. (2014). Estimating population treatment effects from a survey subsample. *American Journal of Epidemiology*, 180(7), 737–748. <https://doi.org/10.1093/aje/kwu197>.
- Sato, T., & Matsuyama, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology*, 14(6), 680–686. <https://doi.org/10.1097/01.EDE.0000081989.82616.7d>.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21. <https://doi.org/10.1214/09-STS313>.

- Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A*, 174(2), 369–386. <https://doi.org/10.1111/j.1467-985X.2010.00673.x>.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2014). Forecasting elections with non-representative polls. *International Journal of Forecasting*. <https://doi.org/10.1016/j.ijforecast.2014.06.001>.
- Zhang, X., Holt, J. B., Yun, S., Lu, H., Greenlund, K. J., & Croft, J. B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology*. <https://doi.org/10.1093/aje/kwv002>.

State Longitudinal Data System

Ting Zhang

Department of Accounting, Finance and Economics, Merrick School of Business, University of Baltimore, Baltimore, MD, USA

Definition

State Longitudinal Data Systems (SLDS) connect databases across two or more of state-level agencies of early learning, K–12, postsecondary, and workforce. It is a state-level Integrated Data System and focuses on tracking individuals longitudinally.

Purpose of the SLDS

SLDS are intended to enhance the ability of states to capture, manage, develop, analyze, and use student education records, to support evidence-based decisions to improve student learning, to facilitate research to increase student achievement and close achievement gaps (National Center for Education Statistics 2010), to address potential recurring impediments to student learning, to measure and document education long-term return on investment, to support education accountability systems, and to simplify the

processes used by state educational agencies to make education data transparent through federal and public reporting (US Department of Education 2015). The Statewide Longitudinal Data Systems Grant Program funds states' efforts to develop and implement these data systems in response to legislative initiatives (US Department of Education 2015).

Information Offered

The data system aligns p-12 student education records with secondary and postsecondary education and the workforce records, with linkable student and teacher identification numbers and student and teacher information on student level (National Center for Education Statistics 2010). The student education records include information on enrollment, demographics, program participation, test records, transcript information, college readiness test scores, successful transition to postsecondary programs, enrollment in postsecondary remedial courses, entries, and exits from various levels of the education system (National Center for Education Statistics 2010).

Statewide Longitudinal Data Systems Grant Program

According to US Department of Education (2015), the Statewide Longitudinal Data Systems Program awards grants to State educational agencies to design, develop, and implement SLDS to efficiently and accurately manage, analyze, disaggregate, and use individual student data. As authorized by the Educational Technical Assistance Act of 2002, Title II of the statute that created the Institute of Education Sciences (IES), the SLDS Grant Program has awarded competitive, cooperative agreement grants to almost all states since 2005; in addition to the grants, the program offers many services and resources to assist education agencies with SLDS-related work (US Department of Education 2016).

Challenges

In addition to the challenges an Integrated Data System has, SLDS has the following main challenges:

Training/Education Provider Participation

In spite of the recent years' progress, participation by training/education providers has not been universal. To improve the training and education coverage, a few states have taken effective action. For example, the Texas state legislature has tied a portion of the funding of state technical colleges to their ability to demonstrate high levels of program completion and employment in occupations related to training (Davis et al. 2014).

Privacy Issues and State Longitudinal Data Systems

To ensure data privacy and protect personal information, Family Educational Rights and Privacy Act (FERPA), the Pupil Protection Rights Act (PPRA), and Children's Online Privacy Protection Act (COPPA) are issued (Parent Coalition for Student Privacy 2017). However, the related issues and rights are complex, and the privacy rights provided by law are often not provided in practice (National Center for Education Statistics 2010). For a sustained SLDS, a push in the established privacy rights is important.

FERPA Interpretation

Another challenge is that some state education agencies have been reluctant to share their education records, largely due to narrow state interpretations of the confidentiality provisions of FERPA and its implementing regulations (Davis et al. 2014). Many states have overcome potential FERPA-related obstacles in their own unique ways, for example: (1) obtaining legal advice recognizing that the promulgation of amended FERPA regulations was intended to facilitate the use of individual-level data for research purposes, (2) maintaining the workforce data within the education state's agency, and (3) creating a special agency that holds both the education and workforce data (Davis et al. 2014).

Maintaining Longitudinal Data

Many state's SLDS already have linked student records, but decision making based on a short-term return on education investment is not necessarily useful; the word "longitudinal" is the keystone needed for development of a strong business case for sustained investment in a SLDS (Stevens and Zhang 2014). "Longitudinal" means the capability to link information about individuals across defined segments and through time. While there is no evidence that the length of data retention increases identity disclosure risk, public concern about data retention is escalating (Stevens and Zhang 2014).

Examples

Examples of US SLDS include:

Florida Education & Training Placement Information Program
 Louisiana Workforce Longitudinal Data System (WLDS)
 Minnesota's iSEEK data.
 Heldrich Center data at Rutgers University
 Ohio State University's workforce longitudinal administrative database
 University of Texas Ray Marshall Center database,
 Virginia Longitudinal Data System
 Washington's Career Bridge
 Connecticut's Preschool through Twenty and Workforce Information Network
 Delaware Education Insight Dashboard
 Georgia Statewide Longitudinal Data System and Georgia Academic and Workforce Analysis and Research Data System (GA AWARDS)
 Illinois Longitudinal Data System
 Indiana Network of Knowledge (INK),
 Maryland Longitudinal Data System
 Missouri Comprehensive Data System
 Ohio Longitudinal Data Archive (OLDA)
 South Carolina Longitudinal Information Center for Education (SLICE)
 Texas Public Education Information Resource (TPEIR) and Texas Education Research Center (ERC)
 Washington P-20W Statewide Longitudinal Data System.

Conclusion

SLDS connects databases across two or more of agencies of p-20 and Workforce. It is a US state-level Integrated Data System and focuses on tracking individuals longitudinally. SLDS are intended to enhance the ability of states to capture, manage, design, develop, analyze, and use student education records and to support data-driven decisions to improve student learning and to facilitate research to increase student achievement and close achievement gaps. The Statewide Longitudinal Data Systems (SLDS) Grant Program funds states' efforts to develop and implement these data systems in response to legislative initiatives. The main challenges of SLDS include training/education provider participation, privacy issues and State Longitudinal Data Systems, and FERPA interpretation, and maintaining longitudinal data. There are many Nationwide SLDS examples.

Cross-References

► [Integrated Data System](#)

Further Reading

- Davis, S., Jacobson, L., & Wandner, S. (2014). *Using workforce data quality initiative databases to develop and improve consumer report card systems*. Washington, DC: Impaq International.
- National Center for Education Statistics. (2010). "Data stewardship: Managing personally identifiable information in student education records." SLDS technical brief. Available at <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011602>.
- Stevens, D., & Zhang, T. (2014). "Toward a business case for sustained investment in State Longitudinal Data Systems." Jacob France Institute. Available at <http://www.jacob-france-institute.org/wp-content/uploads/JFI-WDQI-Year-Three-Research-Report1.pdf>.
- US Department of Education. (2015). "Applications for new awards; Statewide Longitudinal Data Systems Program." Federal register. Available at <https://www.federalregister.gov/documents/2015/03/12/2015-05682/applications-for-new-awards-statewide-longitudinal-data-systems-program>.
- US Department of Education (2016). "Agency information collection activities; Comment request; State Longitudinal Data System (SLDS) Survey 2017–2019."

Federal Register. Available at <https://www.federalregister.gov/documents/2016/10/07/2016-24298/agency-information-collection-activities-comment-request-state-longitudinal-data-system-slids-survey>.

Parent Coalition for Student Privacy (2017). Federal Student Privacy Rights: FERPA, PPRA AND COPPA, retrieved on May 14, 2017 from the World Wide Web https://www.studentprivacymatters.org/ferpa_ppra_coppa/.

Statistician

► [Data Scientist](#)

Statistics

► [“Small” Data](#)

Storage

Christopher Nyamful¹ and Rajeev Agrawal²
¹Department of Computer Systems Technology,
 North Carolina A&T State University,
 Greensboro, NC, USA

²Information Technology Laboratory, US Army
 Engineer Research and Development Center,
 Vicksburg, MS, USA

Introduction

Data storage generally refers to the keeping of data in an electronic or a hard copy form, which can be processed by a computer or a device. Most data today are captured in electronic format, processed, and stored likewise. Data storage is a key component of the Information Technology (IT) infrastructure. Different types of data storage, such as, on-site storage, remote storage, and more recently, cloud storage, play different roles in the computing environment. Huge streams of data are being generated daily. Data activities from social media, data-intensive applications,

scientific research, and industries are increasing exponentially. These huge volumes of data sets must be stored for analytical purposes and also to be compliant with state laws such as the Data Protection Act.

Companies such as YouTube receives one billion unique users each month, and 100 hours of video are uploaded to YouTube every minute (YouTube Data Statistics 2015). Flickr receives on the average 3.5 million uploaded images daily, and Facebook processes 300 million photos per day, and scans roughly 105 terabytes of data each half hour. Storing these massive volumes of data has become problematic, since the conventional data storage reaches a bottleneck. The storage demand for big data at the organizational level is reaching petabytes (PB) and even beyond. A new generation of data storage systems that focuses on large data sets has now become a research focus.

An ideal storage system comprises of a variety of components, including disk arrays, storage controllers, servers, storage network switches, and management software. These key components must fit together to achieve high storage performance. Disks or storage devices are fundamental to every storage system out there. Solid-state drives and hard disk drives are mostly used by organizations as their storage device, with their capacity density expected to increase at a rate of 20% (Fontana et al. 2012). Several attributes such as capacity, data transfer rate, access time, and cost influences the choice of disk for a storage system. Magnetic disc, such as the hard disk drive (HDD) provides huge capacity at a relatively low cost. More HDDs can be added to a storage system to scale to meet the rate of data growth, but are subject to reliability risk, such as overheating, external magnetic faults, and electrical faults. Besides, they have relatively poor input/output operations per second (IOPS) capabilities. Solid-state disks (SSDs) on the other hand, are more recent and more reliable than HDDs. They provide a high aggregate input/output data transfer rate and consume less energy in a storage system. The disadvantage is that they are very expensive per the capacity they provide.

Big data storage systems face complex challenges. Big data has outgrown its current infrastructure, and its complexities translate into variables such as volume, velocity, and variety. Big data means big storage. The demand for storage capacity and scalability has become a huge challenge for large organizations and governmental agencies. The existing traditional system cannot efficiently store and support processing of these data. Data is being transmitted and received from every conceivable direction. To enable high-velocity capture, big data storage system must process with speed. Clients and automated devices demands real-time or near real-time response in order to function or stay in business. Late results from a storage system are of no or little value. In addition to speed, big data comes in different forms. They may consist of structured data – tables, log files and other database files, semi-structured data and unstructured data such as pictures, blogs, and videos. There has to be a connection and correlation between these diverse data types. The complex relationship between this data types cannot be efficiently processed by traditional storage systems.

Storage Systems

Different types of data storage systems serve different purposes. Organizational data requirement usually determines the choice of storage system. Small to medium size enterprises may prefer to keep on-site data storage. Direct-attached storage (DAS) is an example of on-site storage system. DAS architecture connects storage device directly to hosts. This connection can be internal or external. External DAS often attached dedicated storage arrays directly to their host, and data can be accessed at both block level and file level. DAS provides users with enhanced performance than network storage, since host does not have to traverse the network in order to read and write data. Communication between storage arrays and hosts can be over small computer system interface (SCSI) or Fibre Channel (FC) protocol. DAS are easy to deploy and manage. In big data environment, DAS is highly limited in terms of

performance and reliability. DAS can't be shared among multiple nodes, and hence, when one server fails, there is no failover to ensure availability. The storage array has a limited number of ports, making DAS not to scale well to meet the demands of data growth.

The efficient dissemination of mission-critical data among clients over a wide geographical area is very crucial in big data era. Network-attached storage (NAS) infrastructure provides the flexibility of file sharing over a wide area network. NAS achieve this by consolidating wide spread storage used by clients into a single system. It makes use of file sharing protocols to provide access to the storage units. A NAS device can exist anywhere on the local area network (LAN). The device is optimized for cross platform file services such as file sharing, retrieving, and storing. NAS comes with its own operating system, optimized to enhance performance and throughput. It provides a centralized and a simplified storage management to minimize data redundancy on client workstation. For large data sets, a scale-out NAS can be implemented. More storage nodes can be added while maintaining performance and low latency. Despite the functionalities provided by NAS, it still has some shortcomings. NAS operate on the internet protocol (IP) network; therefore, factors such as bandwidth and response time that affects IP networks, equally affects NAS performance. The massive volumes of data can increase latency since IP network can process input/output operations in a timely manner.

Storage area networks (SAN) employs a technology that to a considerable extent deals with the challenges posed by big data storage. SAN architecture comes in two forms – Fibre Channel (FC) SAN and IP-SAN. FC-SAN uses a Fibre Channel protocol to communicate between host and storage devices. Fibre Channel is a high-speed, high-performance network technology that increases data transfer rate between hosts and large storage systems. FC-SAN architecture is made up of servers, FC switches, connectors, storage arrays, and management software. The introduction of FC switches has enhanced the performance of FC-SAN to be highly scalable and to enable better data accessibility. FC-SAN focuses on

consolidating storage nodes to increase scalability, to facilitate a balance input/output operation, and a high performance throughput capability. Implementing FC-SAN architecture is very costly. Besides, it has limited the distance it can go, averagely about 10 km. Large organizations are striving to achieve the best out of their storage systems, while maintaining a low cost. To make use of existing IP-based infrastructure of organizations, IP-SAN technology allows block data to be sent across IP networks. The widespread of IP networks makes IP-SAN attractive to many organizations that are widespread geographically.

More often, big data analysis and processes involves both block-level and file-level data. Object storage technology provisions for file and block-based data storage. Storage object is a logical collection of discrete units of data storage. Each object includes data, metadata, and a unique identifier which allows the retrieval of object data without the need to know the actual location of the storage device. Objects are of variable sizes and ideal to store different types of data that are found in the big data environment. Object-based storage metadata capabilities and flat addressing process allows it to scale with data growth as compared to file system approach. The idea of data and metadata to be stored together ensures easier manageability and migration for long-term storage. Object-based storage is a unified system which combines the advantages of both NAS and SAN. This makes it ideal for storing the massive growth of unstructured data such as photos, videos, tweets, and blogs. It also makes it attractive for cloud deployments.

Active data centers provide data storage service capabilities to clients. It makes use of virtualization to efficiently deploy storage resources to organizations. Virtualization ensures flexibility of resource utilization and optimizes resource management. Data centers are built to host and manage very large data sets. Large organizations keep their data across multiple data centers to ease workload processes and as a backup, in case of any eventualities. The storage layer of a data center usually consists of servers, storage devices, switches, routers, and connectors. Fibre Channel switch is used in a SAN data center for

high transmission of data or commands between servers and storage disk. Storage network devices provide the needed connectivity between hosts to storage nodes. Data center environment supports high-speed transmission of data for both block-level access, supported by SAN, and file-level access, which is also supported by NAS. The consolidation of applications, servers, and storage under one central management increases flexibility and high performance in the big data settings.

Distributed Storage

Distributed file systems, such as Hadoop Distributed File Systems (HDFS) (White 2012), have become very significant in the era of big data. It provides a less expensive but reliable alternative for current data storage systems. It runs on low-cost hardware. HDFS is optimized to run huge volumes of data sets – terabytes and petabytes. It provides high performance data transfer rate and scalability to multiple nodes in a single cluster. HDFS are designed to be very reliable. It stores files as a sequence of blocks of data. Each block of data is replicated to another storage array to ensure reliability and fault tolerant. HDFS cluster has two types of nodes – NameNode and DataNode. The NameNode optimizes the file system namespace and metadata for all files. Storage and retrieval of block-level data is done by the DataNode as per client request or instruction. The retrieved data is sent back to the NameNode with the list of stored block data. Storage vendors such as EMC are releasing *ViPR HDFS Storage* and *Isilon HDFS Storage* for large enterprises and data centers. These systems allow large organizations to roll an HDFS file system over their existing data in place, to perform various services efficiently.

Data backup and recovery has played a significant role in data storage systems. By creating additional copy of production data, organizations are insured against corrupt or deleted data, and recovery of lost data in case of extreme disaster. Besides this objective, there is also the need for compliance with regulatory standards for data storage. The retention of data to ensure high

availability brings data backup, archiving, and replication into the storage domain. Because organizations require quick recovery from backups, IT departments and storage vendors are faced with a huge challenge affecting these activities in big data environment. An ideal backup solution should ensure minimal loss of data, avoid storage of redundant data, and efficient recovery method. The rate of data loss and downtime of an organization in terms of RPO and RTO determines the backup solution to choose. Recovery point in time (RPO) is the point in time from which data must be restored in order to resume processing transactions. RPO determines backup frequencies. Recovery time objective (RTO) is the period of time allowed for recovery, time that can elapse between the disaster and the activation of secondary site. Backup media or storage devices can significantly affect data recovery time, especially with large data sets.

The implementation of large-scale storage system is not straightforward. An ideal storage system comprises a well-balanced components that fits together to achieve optimal performance. Big data requires a high-performance storage system. Such a system usually consists of a cluster of host, interconnected with high-speed network to an array of disks. An example is Lustre file system (LFS), which was partly developed by Hewlett Packard and Intel. LFS capabilities provide up to 512 PB of storage space for one file system. It has a high throughput of about 2 TB/s in a production system. It can contain up to ten million files in a directory, and two billion files in a system. It allows up to 25,000+ clients access in a production system. It provides high availability of data and supports automated failover to meet no-single-point-of failure requirements. EMC ISILON SCALE-OUT NAS also provides a high-performance distribution file system (Rui-Xia and Bo 2012). It consists of nodes of modular hardware arranged in a cluster. Its operating system combines memory, I/Os, CPUs, and disk arrays in a cohesive storage unit as single file system. It provides the capability to address big data challenges by providing multiprotocol file access, dynamic expansion of file system, and high scalability.

Conclusion

Storage is still evolving with advances in technology. The explosive growth rate of data has outgrown its storage capacity. Organizations and businesses are now more concerned about how to efficiently keep and retain all their data. Storage vendors and data center providers are researching into more possible areas of improving storage systems to completely address the overwhelming data size facing the industry.

References

- Fontana, R. E., Hetzler, S. R., & Decad, G. (2012). Technology roadmap comparisons for TAPE, HDD, and NAND flash: Implications for data storage applications. *Magn IEEE Transactions*, 48(5), 1692–1696. <https://doi.org/10.1109/TMAG.2011.2171675>.
- Rui-Xia, Y., & Bo, Y. (2012). *Study of NAS Secure System Base on IBE*. Paper presented at the Industrial Control and Electronics Engineering (ICICEE), 2012 International Conference on.
- White, T. (2012). *Hadoop: The definitive guide*. ISBN: 9781449311520, O'Reilly Media, Inc.
- YouTube Data Statistics. (2015). Retrieved 01–15-2015, 2015, from <http://www.youtube.com/yt/press/statistics.html>.

Storage Media

- [Data Storage](#)

Storage System

- [Data Storage](#)

Stream Reasoning

- [Data Streaming](#)

Structured Data

- [Data Integration](#)

Structured Query Language (SQL)

Joshua Lee

Schar School of Policy and Government, George Mason University, Fairfax, VA, USA

Introduction

Storing, modifying, and retrieving data is one of the most important tasks in modern computing. Computers will always need to retain certain information for later use, and that information needs to be organized, secure, and easily accessible. For smaller, simpler datasets, applications like Microsoft Excel can suffice for data management. For smaller data transfer needs, *XML* functions effectively. However, when the size and complexity of the information becomes great enough, a complex relational database management system (RDBMS) such as SQL becomes necessary.

SQL stands for *structured query language*. It was initially invented in 1974 by IBM under the name SEQUEL, and was bought by Oracle in 1979, and since then it has become the dominant force in RDBMS (See http://docs.oracle.com/cd/B12037_01/server.101/b10759/intro001.htm).

Unlike programming languages such as C or BASIC, which are *imperative programming languages*, SQL is a *declarative programming language*. The difference is explored in greater detail below.

In a SQL environment, you create one or more *databases*. Each database contains any number of *tables*. Each table in turn contains any number of *rows* of data, and each row contains *cells*, just like an Excel spreadsheet.

Definitions

Declarative programming language	A programming language which expresses the logic of a computation without describing its flow control. SQL is an example of a declarative programming language: SQL commands describe the <i>logic</i> of computation in great detail, but SQL does not contain flow control elements (such as IF statements) without extensions.
Imperative programming language	A programming language which focuses on <i>how</i> a program should operate, generally via flow-control statements. C and BASIC are examples of imperative programming languages.
XML	Stands for Extensible <i>markup language</i> . XML documents are designed to store data in such a way as it is easily readable both for human beings and computers. XML is a common format for sending organized, structured data across the Internet.
Flat file database	A database in which there is a single table. Microsoft Excel spreadsheets are a popular example of a flat file database.
Relational database management system (RDBMS)	A database which contains multiple tables that are related (linked) to one another in various possible ways.
Table	A set of data values represented by rows and columns
SQL statement	Shorthand for the data manipulation language category of SQL commands. See Types of SQL Commands below for more information.
SQL query	Shorthand for the data query language category of SQL commands. See Types of SQL

	Commands below for more information.
SQL clause	The building blocks of both SQL statements and queries.

Flat File Databases Versus Relational Databases: A Comparison

Importantly, RDBMS are not always the correct solution. While they allow for a much greater degree of control, optimization in processing speed, and handling of complex relational data, they also require substantially more time and skill to configure and maintain properly. Therefore, for simpler, non-relational data, there is no problem with using a flat file database.

Visualization is a critical component of understanding how relational databases (such as SQL) compare to flat file databases. Consider the following Excel spreadsheet:

Orders		
Order_No	Name	City
46321	John Doe	Los Angeles
94812	James Hill	Miami
29831	Maria Lee	Austin
59822	James Hill	Miami

This example represents a classic database need: there is an online business, and that business needs to store data. They naturally want to store the orders that each of their customers have made for later analysis. Thus, there are two kinds of data being represented in a single spreadsheet: the orders that a customer makes and the customer’s information. These two types of information are distinct, but they’re also inherently related to one another: orders only exist because a specific customer makes them.

However, storing this information as a flat file database is flawed. First and foremost, there is a *many-to-one relationship* at play: one person can have multiple orders. To represent this in an Excel spreadsheet, you would thus need to have

multiple rows dedicated to a single user, with each row differing simply in the order that the customer made. We can see this in the second and fourth *Orders* rows: James Hill has his information repeated for two separate orders. This is a highly inefficient method of data storage. In addition, it causes data modification tasks to become more complicated and time-consuming than necessary.

For example, if a customer needed to change their City, the data for every single order that person has made must be modified. This is needlessly repetitive – the customer’s

information needs to be repeated for every order, even though it should always be identical. Now, instead of the small amount of data above, imagine a spreadsheet with 10,000 orders and 2000 unique customers. Clearly, the amount of extraneous data needed to represent this information is huge – with an average of five orders per person, that person’s name and city must be repeatedly entered. For this problem, a flat file database has both excessive file size and excessive processing time.

Next, consider how a RDBMS such as SQL would handle this situation:

Online_Store_DB

Persons		
Person_ID	Name	City
1	James Hill	Miami
2	John Doe	Los Angeles
3	Maria Lee	Austin

Orders		
Order_ID	Order_No	Person_ID
1	46321	2
2	94812	1
3	29831	3
4	59822	1

This SQL database is named *Online_Store_DB*, and it contains two tables: *Persons* and *Orders*. *Persons* contains three rows and three columns, and *Orders* contains four rows and three columns. The first immediately noticeable difference is that SQL splits this dataset into its two distinct tables – a Person and an Order are two different things, after all. However, even if you split an Excel spreadsheet in the same way, SQL does more than merely split the data. It is the *relational* aspect of SQL that allows for a far more elegant solution.

Looking above, both the *Persons* and the *Orders* tables have a column titled *Person_ID*. Each Person will always get their own *Person_ID* in the *Persons* table, so it’s always unique. What’s more, the *Orders* table is *linked* to the *Persons* table through that column (in SQL parlance, *Person_ID* in the *Orders* table is known as a *foreign key*). Rather than repeat all the

information (as the flat file database does), each Order is linked to their associated customer’s *Person_ID*. Thus, by modifying the *City* in a single row in the *Persons* table, all *Orders* associated with that person both (a) don’t need to be changed and (b) will always be linked to the most current information.

Notably, even if you attempted to create such a structure in two different Excel spreadsheets, it still wouldn’t be able to achieve the same result. Excel doesn’t allow for this kind of linking of columns in different spreadsheets. This linking not only speeds up the querying of data substantially; it also adds fail-safe features to ensure that modifications to one table don’t inadvertently corrupt data in another table. For example, SQL can (optionally) stop you from deleting a person from the *Persons* table if they currently have any *Orders* associated with them – after all, orders shouldn’t exist without an associated person. By contrast,

it's easy to imagine two (or more) different Excel spreadsheets having inconsistent data appear over a long period.

Types of SQL Commands

SQL functions via the use of *commands*. These commands can be separated into six general categories: DDL (data definition language), DML (data manipulation language), DQL (data query language), DCL (data control language), data administration commands, and transactional control commands.

- DDL includes all SQL commands to create/modify/delete entire databases or tables, such as CREATE, ALTER, and DROP.
- DML includes all SQL commands to manipulate all data stored *inside* of tables, such as INSERT, UPDATE, and DELETE, among others.
- DQL includes only a single SQL command, SELECT, which focuses exclusively on retrieving data from within SQL. Rather than creating/modifying/deleting data, DQL focuses simply on grabbing the data for use by the user. SELECT statements can have many possible *clauses*, such as WHERE, GROUP BY, HAVING, ORDER BY, JOIN, and AS. These clauses modify how the SELECT statement functions.
- DCL includes all SQL commands to control user permissions within SQL. For example, if User A should be able to view Table A and Table B, but User B should only be allowed to view Table A. Additionally, User B should not be permitted to modify any data he views. DCL commands include ALTER PASSWORD, REVOKE, and GRANT, among others.
- Data administration controls focus on analyzing the performance of other SQL commands – how quickly are they processed, how often are certain SQL queries used, and where are the greatest bottlenecks in performance. AUDIT is a common data administration control command.
- Transactional control commands are for controlling whether certain SQL queries *should* be executed. For example, perhaps the user needs three consecutive SQL queries to be run, Queries A, B, and C. However, the user also wants to ensure that if *any one* of the three queries fails to execute properly, the other two queries are immediately reversed and the database is restored to its previous state. Transactional control commands include COMMIT, ROLLBACK, and SAVEPOINT, among others.

SQL Extensions

Multiple organizations have created their own unique “extended” versions of SQL with additional capabilities. Some of these versions are public and open-source, whereas others are proprietary. While these extended versions of SQL generally don't remove any of its core capabilities as described above, they instead add features on top. These additional capabilities usually involve flow control.

For example, one popular example of SQL with such an extension is MySQL (See <https://dev.mysql.com/doc/refman/5.7/en/forofficialdocumentation>). MySQL is open-source and has all the capabilities of SQL noted above. However, it also provides SQL with flow control capabilities like an imperative programming language. The full list of such features is beyond the scope of this section, but some of the most important include (1) stored procedures, (2) triggers, and (3) nested SELECT statements. Notably, these three features are also very common in other extensions of SQL.

Stored procedures are specific collections of SQL commands that can be run through with a single EXECUTE command. Not only can multiple different SQL commands be executed in order, but they also can be run using common flow-control mechanisms found in any imperative programming language. These include IF statements, WHILE loops, RETURN values, and iterating through values obtained via SELECT statements as if they were an array. It also includes the ability

to store values temporarily in variables outside of the permanent database. While much of this logic could theoretically be handled by another programming language that connects to SQL, executing this logic directly in SQL can increase both performance and security since the processing is all done server-side (See <https://www.sitepoint.com/stored-procedures-mysql-php/formoreinformation>).

Triggers are related to stored procedures in that they are also collections of SQL commands. However, whereas stored procedures are generally activated by a user, triggers are event-based in their activation. For example, a trigger can be set to run every 30 min, or can be set to run every X queries that are executed, or whenever a certain table is modified.

Nested SELECT statements are another feature of MySQL. They allow for SELECT statements to contain other SELECT statements inside of them. This nesting allows for on-the-fly sorting and filtering of complex data without needing to make unnecessary database modifications along the way.

Further Reading

Introductory Book: SQL in 10 Minutes, Sams Teach Yourself (4th edn). ISBN: 978-0672336072

Online Interactive SQL Tutorial: <https://www.khanacademy.org/computing/computer-programming/sql/sql-basics/v/welcome-to-sql>.

Quick-Start SQL Command Cheat Sheet: https://www.w3schools.com/sql/sql_intro.asp.

Stylistics

- [Authorship Analysis and Attribution](#)

Stylometry

- [Authorship Analysis and Attribution](#)

Supercomputing, Exascale Computing, High Performance Computing

Anamaria Berea

Department of Computational and Data Sciences,
George Mason University, Fairfax, VA, USA
Center for Complexity in Business, University of
Maryland, College Park, MD, USA

Supercomputing and High Performance Computing (HPC)

Supercomputing and High Performance Computing are synonymous; both terms reflect a computational system that is measured in FLOPS and which requires a complex computing architecture. On another hand, Exascale computing is a specific type of super computing, with a computational power of one billion calculations per second. While there are supercomputing and high performance computing systems in existence now in various countries (the USA, China, India, and the EU), the Exascale computing has not been achieved yet.

Supercomputing is proved to be very useful for large-scale computational models, such as weather and climate change models, nuclear weapons and security simulations, brute force decryption, molecular dynamics, the Big Bang and the beginning of the Universe, gene interactions, and simulations of the brain.

Supercomputers will represent significant human capital and innovation. The USA offers current opportunities for HPC access to any American company that demonstrates strategies to “make the country more competitive.” High Performance Computing will emerge as the ultimate signifier of talent and scientific prestige; at least one study found that universities that invest in supercomputers have a competitive edge in research. Meanwhile, Microsoft has reorganized HPC efforts into a new “big compute” team, denoting a new era of supercomputing.

On another hand, there are many challenges that come with achieving supercomputing and

HPC, among which there are the “end of Moore’s Law,” parallelization mechanisms, and economic costs. Specifically, current research efforts into supercomputing are not focused on improving clock speeds as in classic Moore’s Law, but on improving speeds to core and parallelization. On another hand, parallelization efforts in supercomputing are not focused on improving the system design, but the architecture design. And lastly, these efforts come at a significant price: currently, the USA has invested more than \$120 million into supercomputing and high performance computing research.

Some other important challenges for Exascale and HPC computing in terms of technology are resilience and scalability; with the physical increase in the systems there also comes a decrease in resilience and their scalability is more difficult to achieve (Shalf et al. 2010).

Supercomputing Projects Around the World

The European Commission estimates that High Performance Computing (HPC) will accelerate the speed of big data analysis toward a future where a variety of scientific, environmental, and social challenges can be addressed, especially on extremely large and small scales (IDC 2014). Tens of thousands of times more powerful than laptop computers, HPC conducted on supercomputers processes information using parallel computing, allowing for many simultaneous computations to occur at once. These integrated machines are measured in “flops” which stands for “floating point operations per second.”

As of June 2013, Tianhe-2 (in translation Milky Way-2), a supercomputer developed by China’s National University of Defense Technology is the world’s fastest system with a performance of 33.86 petaflop/s.

Exascale Computing

During the past 20 years, we have witnessed the move from terascale to petascale computing. For

example, Pleiades was NASA’s first petascale computer (Vetter 2013).

The current forecasts place the HPC into “Exascale” capacity by 2020, developing computing capacities 50 times greater than today’s most advanced supercomputers. Exascale feasibility depends on the rise of energy-efficient technology: the processing power exists but the energy to run it, and cool it, does not. Currently, the American supercomputer MIRA, while not the fastest, is the most energy efficient, thanks to circulating water-chilled air around the processors inside the machine rather than merely using fans.

Intel has revealed successful test results on servers submerged in mineral oil liquid coolant. Immersion cooling will impact the design, housing, and storage of servers and motherboards, shifting the paradigm from traditional air cooling to liquid cooling and increasing the energy efficiency of HPC. Assurances that computer equipment can be designed to withstand liquid immersion will be important to the development of this future. Another strategy to keep energy costs down is ultra-low-power mobile chips.

Applications of HPC

Promising applications of HPC to address numerous global challenges exist:

- Nuclear weapons and deterrence: monitors the health of America’s atomic arsenal and performs “uncertainty quantification” calculations to pinpoint the degree of confidence in each prediction of weapons behavior.
- Supersonic noise: a study at Stanford University to better understand impacts of engine noise from supersonic jets.
- Climate change: understanding natural disasters and environmental threats, predicting extreme weather.
- Surveillance: intelligence analysis software to rapidly scan video images. IARPA (Intelligence Advanced Research Projects Activity) requests the HPC industry to develop a small computer especially designed to address intelligence gathering and analysis.

- Natural Resources: simulating the impact of renewable energy on the grid without disrupting existing utilities, making fossil fuel use more efficient through modeling and simulations of small (molecular) and large-scale power plants.
- Neuroscience: mapping the neural structure of the brain.

The HPC future potentially offers solutions to a wide range of seemingly insurmountable critical challenges, like climate change and natural resource depletion. The simplification of “big problems” using “big data” with the processing power of “big compute” runs the risk of putting major decisions at the mercy of computer science. Meanwhile, some research has suggested that crowdsourcing (human data capture and analysis) may in fact exceed or improve outcomes of supercomputer tasks.

Also in the USA, the Oak Ridge National Laboratory (ORNL) provides access to the scientific research community to the largest US supercomputer. Some of the projects that are using the Oak Ridge Leadership Computing Facility’s supercomputer are in the fields of sustainable energy solutions, nuclear energy systems, advanced materials, sustainable transportation, climate change, and the atomic-level structure and dynamics of materials.

The Exascale Initiative launched by the Department of Energy, National Nuclear Security Administration (NNSA), and the Office of Science (SC) has a main goal to “target the R&D, product development, integration, and delivery of at least two Exascale computing systems for delivery in 2023.”

Supercomputing and Big Data

The link between HPC or Exascale computing and big data is obvious and intuitive in theory, but in practice this is more difficult to achieve (Reed and Dongarra 2015). Some of these implementation challenges come from the difficulty of creating resilient and scalable data architectures.

On another hand, the emergence of GPUs will likely help solve some of these current challenges (Keckler et al. 2011). GPUs are particularly good at handling big data.

Some supercomputers can cost even up to \$20 million, and they are made of thousands of processors. Alternatively, clusters of computers work together as a supercomputer. For example, a small business could have a supercomputer with as few as 4 nodes, or 16 cores. A common cluster size in many businesses is between 16 and 64 nodes, or from 64 to 256 cores.

As power consumption of supercomputers increases though, so does the energy consumption necessary to cool and maintain the physical infrastructure of HPCs. This means that energy efficiency will move from desirable to mandatory and currently there is research underway to understand how green HPC or efficient energy computing can be achieved (Hemmert 2010).

Architectures and Software

The architectures of HPC can be categorized in: (1) commodity-based clusters, with standard HPC software stacks, from Intel or AMD; (2) GPU-accelerated commodity-based clusters (GPUs are specific for gaming and professional graphics markets); (3) customized architectures, with customization both for the nodes and for their interconnectivity networks (i.e., K computer and Blue Gene systems); and (4) specialized systems (i.e., protein folding). These last systems are more robust, but less adaptable, while the customized ones are the most adaptable to efficiency, resilience, and energy consumption (Vetter 2013).

HPC systems share many data architectures and servers, but the software is much more integrated and hierarchical. The HPC software stack has a system software, a development software, a system management software, and a scientific data management and visualization software. These include operating systems, runtime systems, and file systems. These facilitate programming models, compilers, scientific frameworks and libraries, and performance tools.

Outside of HPC, clouds and grids have increased in popularity (i.e., Amazon EC2). These are specific for data centers and enterprise markets, as well as internal corporate clouds.

Further Reading

- Hemmert, S. (2010). Green HPC: From nice to necessity. *Computing in Science & Engineering*, 12(6), 8–10.
- IDC. (2014). *High performance computing in the EU: Progress on the implementation of the European HPC strategy*. Brussels: European Commission., ISBN: 978-92-79-49475-8. <https://doi.org/10.2759/034719>.
- Keckler, S. W., et al. (2011). GPUs and the future of parallel computing. *IEEE Micro*, 31(5), 7–17.
- Reed, D. A., & Dongarra, J. (2015). Exascale computing and big data. *Communications of the ACM*, 58(7), 56–68.
- Shalf, J., Dosanjh, S., & Morrison, J. (2010). Exascale computing technology challenges. In *International conference on high performance computing for computational science*. Berlin/Heidelberg: Springer.
- Vetter, J. S. (Ed.). (2013). *Contemporary high performance computing: From Petascale toward Exascale*. Boca Raton: CRC Press.

environment and affecting supply chain management practices (Min et al. 2019) as firms look for opportunities to improve their long-term performance. While supply chain management has always been technology-oriented and data-intensive, the ongoing explosion of big data, and the tools to make use of this data, is opening many avenues to advance decision-making along the supply chain (Alicke et al. 2019). Big data enables companies to use new data sources and analytical techniques to design and run smarter, cheaper, and more flexible supply chains. These benefits can often be observed in areas making use of automated, high-frequency decisions, such as demand forecasting, inventory planning, picking, or routing. However, also other supply chain activities benefit from the different Vs of big data (e.g., volume, variety, or velocity). This improved decision-making, often supported by artificial intelligence (AI) and machine learning (ML) approaches, is frequently referred to as supply chain analytics.

Supply Chain Activities

Supply chain management encompasses many complex decisions along the various supply chain activities. Here, we distinguish between strategic, tactical, and operational decisions. Strategic decisions, such as network design or product design, typically have a long-term impact. As such, they rely on a holistic perspective that requires data as an input along with human judgment. For planning and execution with a mid- to short-term focus, big data offers tremendous possibilities to improve key decisions.

Figure 1 outlines the key decisions along the supply chain for a typical consumer goods product. The complexity of supply chain operations is obvious: raw materials are sourced from hundreds of suppliers for thousands of stock keeping units (SKUs) that are produced in many plants worldwide and are delivered to tens of thousands of stores. The overarching *sales and operations planning* (S&OP) process is centered around demand, production, and inventory planning and aims to create a consensus, efficient plan that

Supply Chain and Big Data

Kai Hoberg
Kühne Logistics University, Hamburg, Germany

Introduction

Supply chain management focuses on managing the different end-to-end flows (i.e., material, information, and financial flows) within a particular company and across businesses within the supply chain (Min et al. 2019). As such, it encompasses all activities along the value chain, e.g., from planning, sourcing, and manufacturing to warehousing and transportation, in collaboration with suppliers, customers, and third parties. Many of these activities require different business functions and cross-functional teams for successful decision-making. In recent years, new and existing technologies have been introduced that are dramatically changing the business



Supply Chain and Big Data, Fig. 1 Mid- and short-term decisions in key supply chain activities

aligns supply and demand. These activities leverage data to minimize forecast errors, improve production plan stability, and minimize inventory holding and shortage costs. In *sourcing*, activities are focused on supplier selection decisions, managing the risk of supply chain disruptions and shortages, and continued cost improvement. Among many other purposes, data is used to model product and supplier cost structures, to forecast disruptions from suppliers at different tiers, and to track and manage supplier performance. In *manufacturing*, activities are centered around detailed production scheduling, quality control, and maintenance. Here, the key objectives typically include improving the overall asset effectiveness (OEE) of production equipment, reducing maintenance costs, and diagnosing and improving processes. Next, in *warehousing*, activities are centered around storing, picking, and packing, often including some additional value-added services (VAS). Data is leveraged, e.g., to allocate goods to storage areas, to reduce walking distances, to increase process quality, and to redesign processes. Further, in *transportation*, mid- to short-term activities are centered around asset management (for trucks, ships, or planes), allocating transportation

jobs to assets, loading goods, and defining truck routing based on ETA (estimated time of arrival). Using the available data, transportation times and costs can be reduced, e.g., by optimizing the routing according to the customer preferences, road conditions, and traffic. Finally, goods are handled at the *point of sale* at the retailer (or point of consumption for other industrial settings). Here, activities are centered around shelf-space optimization, inventory control, detecting stock outs, and optimizing pricing. Data is used to maximize revenues, reduce lost sales, and avoid waste. Based on the goods handled in the supply chain, activities could have a very different emphasis, and additional key activities, such as returns or recycling, must be carefully considered.

Data Sources

Until recently, enterprise resource planning (ERP) systems (provided by commercial vendors, such as SAP, Oracle, or Sage) have been the primary source of data for supply chain decision-making. Typically, data from ERP systems is sufficiently structured and available at decent quality levels. ERP data includes, among lots of other

information, master data, historic sales and production data, customer and supplier orders, and cost information. However, many additional data sources can be leveraged to enrich the various supply-chain decisions. Data collection plays an essential role as without an efficient and effective approach for capturing data, it would be impossible to carry out data-based analytics (Zhong et al. 2016). Among the many additional data sources, the following set has particular relevance for supply chain management since the acquired information can be leveraged for many purposes.

Advanced Barcodes and RFID Tags

Classic barcodes have long been applied to uniquely identify products across supply chain partners. With the emergence of advanced barcodes (e.g., 2D data matrix or QR codes), additional information, such as batch information, serial number, or expiry date can be stored. In contrast to cheaper barcodes, RFID tags can provide similar information without the need for a direct line of sight. The pharmaceutical industry and fashion industry are among the early adopters of these technologies due to the benefits and their needs for increased visibility.

IoT Devices

The introduction of relatively cheap sensors with the Internet-of-things (IoT) connectivity has triggered numerous opportunities to obtain data in the supply chain (Ben-Daya et al. 2019). Applications include temperature sensors tracking the performance of reefer containers, gyroscopes to monitor shocks during the transportation of fragile goods, voltmeters to monitor the performance of electric engines and to trigger preventive maintenance, GPS sensors to track the routes of delivery trucks, and consumption sensors to track the use of, e.g., coffee machines (Hoberg and Herdmann 2018).

Cameras and Computer Vision

HD cameras are already frequently applied and enhance the visibility and security in the supply chain. For example, cameras installed in production lines measure the product quality and automatically trigger alerts based on deviations. In

retail stores, camera-equipped service robots are used to measure inventory levels on store shelves, or fixed cameras can be used to identify customers waiting for assistance and to notify staff to help (Musalem et al. 2016).

Wearables

Easy-to-use interfaces introduced by wearables (e.g., handheld radio-frequency devices, smart glasses) offer location-based and augmented reality-enabled instructions to workers (Robson et al. 2016). However, wearables also offer the potential to log data as the worker uses them. For example, in-door identifiers can track walking paths within plants, barometers can measure the altitude of a delivery in a high-rise building, and eye trackers can capture information about picking processes in warehouses.

Data Streams and Archives

While most analyses focus on historic demand data to forecast future sales, numerous exogenous factors affect demand, and so incorporating them can increase forecast accuracy. Data sources that can be leveraged include macro-level developments (e.g., interest rates, business climate, construction volumes), prices (e.g., competitor prices, market prices, commodity prices), or customer market developments (e.g., motor vehicle production, industrial production). Demand forecasting models can be customized using data streams and archival data relevant to the specific industry or by incorporating company-specific factors, such as the daily weather (Steinker et al. 2017).

Internet and Social Media

Substantial amounts of information are available online and on social media. For example, firms can obtain insights into consumer sentiments and behaviors, crises, or natural disasters in real time. Social media information from Facebook can improve the accuracy of daily sales forecasts (Cui et al. 2018). In a supply chain context, any timely information allows securing an alternative supply in the case of supply chain disruptions or improving sourcing volumes for high-demand items.

In assessing the suitability of the different data sources, it is important to answer several key questions:

- Is the required data directly available to the relevant decision-maker or is there a need to obtain data from other functions or external partners (e.g., promotion information from sales functions for manufacturing decisions, point-of-sale data from retailers for consumer goods manufacturers' demand forecasting)?
- Is the data sufficiently structured to support the analysis or is a lot of pre-processing required (e.g., unstructured customer feedback on delivery performance may require text analytics, data integration from different supply chain partners may be challenging without common identifiers)?
- Is the data sufficiently timely to support the decision-making for the considered activity (e.g., real-time traffic data for routing decisions or weekly information about a supplier's inventory level)?
- Is the data quality sufficient for the purpose of the decision-making? (e.g., is the inventory accuracy at a retail store or master data on the supplier lead times sufficient for replenishment decisions)?
- Is the data volume manageable for the decision-maker, or does the amount of data require support by data engineers/scientists (e.g., aggregate monthly demand data for supplier selection decisions vs. location-SKU-level information on hourly demand for pharmacy replenishment)?

Supply Chain Opportunities for Big Data

The benefits of applying big data for supply chain analytics are generally obvious but often very context-specific (Waller and Fawcett 2013). Many opportunities exist along the different key supply chain processes.

Sales and Operations Planning

Planning can particularly benefit from big data to increase demand forecast accuracy (e.g., using

archives, real-time data feeds, consumption, and point-of-consumption (POC) inventory data) and end-to-end planning. Leveraging concurrent, end-to-end information, the visibility in planning can extend beyond the currently still predominant intra-firm focus (upon partner approval). As a result, detailed information about a supplier's production volumes, in-transit goods, and estimated time of arrivals (ETAs) can reduce costly production changes/express shipments and allow for demand shaping.

Sourcing

In sourcing, big data enables complex cost models to be developed that improve the understanding of cost drivers and optimal sourcing strategies. Further supply chain risk management can benefit from social media information on supply chain disruptions (e.g., accidents, strikes, bankruptcies) as second-, third-, and fourth-tier suppliers are better mapped. Finally, contract compliance can be improved by analyzing shipping and invoice information in real time.

Manufacturing

One way of using big data in manufacturing is to improve the product quality and to increase yields. Information on tolerances, temperatures, or pressures obtained in real time can enable engineers to continuously optimize production processes. Unique part identifiers allow autonomous live corrections for higher tolerances in later manufacturing steps. IoT sensor data enables condition-based maintenance strategies to reduce breakdowns and production downtimes.

Warehousing

Large amounts of data are already widely used in warehousing to increase operational efficiency for storing, picking, and packing processes. Further advances would be possible due to improved forecasts on individual item pick probabilities (to decide on storage location) or by the better prediction of individual picker preferences (to customize recommendations on routing or packing). As warehouses install more goods-to-man butler systems and pickings robots, data is also required

to coordinate machines and to enhance computer vision.

Transportation

To boost long-range transportation and last-mile efficiency, accuracy of detailed future volumes (in air cargo, also the weight) is important. Better data on travel times and for ETA projection (e.g., using weather and traffic data) could further boost asset effectiveness and customer satisfaction. Further, data allows analyzing and managing driver behavior for increased performance and sustainability. Customer-specific information can further improve efficiency, e.g., what are the likely demands that can be bundled, when is the customer likely to be at home, or where to park and how long to walk to the customer's door?

Point of Sale

Bricks-and-mortar stores are increasingly exploring many new data-rich technologies to compete with online competitors. In particular, the potential of data has been recognized to tweak processes and store layouts, to increase customer intimacy for individual advertising, and to advance inventory control and pricing by forecasting individual sales by store. Using IoT sensors and HD cameras, experts can build offline "click-stream" data to track customers throughout the store. Integrated data from online and offline sources can be used to create coupons and customer-specific offers. Finally, accurate real-time stock information can improve inventory control and mark-down pricing.

Conclusion

Big data in supply chain management offers many interesting opportunities to improve decision-making, yet supply chain managers still need to adjust to the new prospects (Alicke et al. 2019; Hazen et al. 2018). However, the ultimate question arises as to if and where the expected value justifies the effort for collecting, cleaning, and analyzing the data. For example, a relatively high increase in forecasting accuracy for a

made-to-order product would not provide tangible benefits, whereas any small increase in ETA accuracy in last-mile delivery can increase routing and improve customer satisfaction significantly. More research is evolving to obtain insights about where big data can provide the most value in supply chain management.

References

- Alicke, K., Hoberg, K., & Rachor, J. (2019). The supply chain planner of the future. *Supply Chain Management Review*, 23, 40–47.
- Ben-Daya, M., Hassini, E., & Bahroun, Z. (2019). Internet of things and supply chain management: A literature review. *International Journal of Production Research*, 57(15–16), 4719–4742.
- Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, 27, 1749–1769.
- Hazen, B. T., Skipper, J. B., Boone, C. A., & Hill, R. R. (2018). Back in business: Operations research in support of big data analytics for operations and supply chain management. *Annals of Operations Research*, 270, 201–211.
- Hoberg, K., & Herdmann, C. (2018). Get smart (about replenishment). *Supply Chain Management Review*, 22(1), 12–19.
- Min, S., Zacharia, Z. G., & Smith, C. D. (2019). Defining supply chain management: In the past, present, and future. *Journal of Business Logistics*, 40, 44–55.
- Musalem, A., Olivares, M., & Schilkret, A. (2016). Retail in high definition: Monitoring customer assistance through video analytics (February 10, 2016). Columbia Business School Research Paper No. 15–73. Available at SRN: <https://ssrn.com/abstract=2648334> or <https://doi.org/10.2139/ssrn.2648334>.
- Robson, K., Pitt, L. F., Kietzmann, J., & APC Forum. (2016). Extending business values through wearables. *MIS Quarterly Executive*, 15(2), 167–177.
- Steinker, S., Hoberg, K., & Thonemann, U. W. (2017). The value of weather information for E-commerce operations. *Production and Operations Management*, 26(10), 1854–1874.
- Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, 34, 77–84.
- Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan S. (2016). Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering*, 101, 572–591.

Surface Web

► Surface Web vs Deep Web vs Dark Web

Surface Web vs Deep Web vs Dark Web

Esther Mead¹ and Nitin Agarwal²

¹Department of Information Science, University of Arkansas Little Rock, Little Rock, AR, USA

²University of Arkansas Little Rock, Little Rock, AR, USA

Synonyms

Dark web; Darknet; Deep web; Indexed web; Indexable web; Invisible web; Hidden web; Lightnet; Surface web; Visible web

Key Points

The World Wide Web (WWW) is a collection of billions of web pages connected by hyperlinks. A minority of these pages are “discoverable” due to their having been “indexed” by search engines, such as Google, rendering them visible to the general public. Traditional search engines only see about 0.03% of the existing web pages (Weimann 2016). This indexed portion is called the “surface web.” The remaining unindexed portion of the web pages, which comprises the majority, is called the “deep web,” which search engines cannot find, but can be accessed through direct Uniform Resource Locators (URLs) or IP (Internet Protocol) addresses. The deep web is estimated to be 400–500 times larger than the surface web (Weimann 2016). Common examples of the deep web include email, online banking, and Netflix, which can only be accessed past their main public page when a user creates an account and utilizes login credentials. There are parts of the deep web called the “dark web” that can only

be accessed by using special technologies, such as the TOR browser, which access the “overlay networks” within which the dark web pages reside (Weimann 2016; Chertoff 2017). Most of the dark web pages are hosted anonymously and are encrypted. Dark web pages are intentionally hidden due to their tendency to involve content of an illegal nature, such as the purchasing and selling of pornography, drugs, and stolen consumer financial and identity information. An effective way to visualize these three web components is as an iceberg in the ocean: Only about 10% of the iceberg is visible above the water’s surface, while the remaining 90% lies hidden beneath the water’s surface (Chertoff 2017).

Technological Fundamentals

The surface web is comprised of billions of statically linked HTML (Hypertext Markup Language) web pages, which are stored as files in searchable databases on web servers that are accessed over HTTP (Hypertext Transfer Protocol) using a web browser application (i.e., Google Chrome, Firefox, Safari, Internet Explorer, Microsoft Edge, etc.). Users can find these surface web pages by using a standard search engine – Google, Bing, Yahoo, etc., – entering keywords or phrases to initiate a search. These web pages can be found because these search engines create a composite index by crawling the surface web, traveling through the static web pages via their hyperlinks. Deep web pages, on the other hand, cannot be found using traditional search engines because they cannot “see” them due to their readability, dynamic nature, or proprietary content; hence the other often-used terms “hidden web” and “invisible web.” The term “invisible web” was coined in 1994 by Jill Ellsworth, and the term “deep web” was coined in 2001 by M. K. Bergman (Weimann 2016). The issue of readability has to do with file types; the issue of dynamic nature has to do with consistent content updates; and the issue of proprietary content has to do with the idea of freemium or pay-for-use sites that require registration with a username and password. Deep web

pages do not have static URL links; consequently, information on the deep web can only be accessed by executing a dynamic query via the query interface of a database or by logging in to an account with a username and password. The dark web (or darknet) is a collection of networks and technologies that operate within a protocol layer that sits on the conventional internet. The term “darknet” was coined in the 1970s to indicate it was insulated from ARPANET, the network created by the U.S. Advanced Research Projects Agency that became the basis for the Surface Web back in 1967. The dark web pages cannot be found or accessed by traditional means due to one or more of the following: (1) The network host is not using the standard router configuration (Border Gateway Protocol); (2) The host server’s IP address does not have a standard DNS entry point because it is not allocated; (3) The host server has been set up to not respond to ping by the Intelligent Contact Manager; and (4) Fast-fluxing DNS techniques were employed that enables the host server’s IP address to continually and quickly change (Biddle et al. 2002). The technologies of the dark web ensure that users’ IP addresses are disguised and their identity remains anonymous by using a series of computers to route users’ activity through so that the traffic cannot be traced. Once users are on the dark web, they can use directories such as the “Hidden Wiki” to help find sites by category. Otherwise, to access a particular dark website, a user must know the direct URL and use the same encryption tool as the site (Weimann 2016).

Key Applications

Key applications for accessing the surface web are common web browsers such as Google Chrome, Firefox, Safari, Internet Explorer, Microsoft Edge, etc. For finding surface web pages, users can use standard search engine applications such as Google, Bing, Yahoo, etc., to enter in keywords or phrases to initiate a search. Accessing the deep web requires database interfaces and queries. There have been some commercial products directed toward the area of trying to enable

searching the deep web including BrightPlanet Deep Query Manager (DQM2), Quigo Technologies’ Intellisionar, and Deep Web Technologies’ Distributed Explorit. Another category of key applications for accessing portions of the deep web are academic libraries that require users to create an account and log in with a username and password; some of these libraries are cost-based if you are not affiliated with an educational institution. Common social media networks such as Facebook, Twitter, and Snapchat are classified as deep web applications because they can only be fully utilized if a user accesses them through their respective application program interface and sets up an account. Popular instant messaging applications such as iChat, WhatsApp, and Facebook Messenger are also part of the deep web, as well as are some file-sharing and storage applications such as DropBox and Google Drive (Chertoff 2017). Key applications for accessing the dark web include peer-to-peer file sharing programs such as Napster, LimeWire, and BitTorrent, and peer-to-peer networking applications such as Tor. Many of these applications have been forced to cease operations, but various subsequent applications continue to be developed (Biddle et al. 2002). Tor is actually a network and a browser application that enables users to browse the internet anonymously by either installing the client or using “a web proxy to access the Tor network that uses the pseudo-top level domain .onion that can only be resolved through Tor” (Chertoff 2017). The Tor network was first presented in 2002 by the United States Naval Research Laboratory as “The Onion Routing project” intended to be a method for anonymous online communication (Weimann 2016; Gehl 2016; Gadek et al. 2018). Using Tor allows users to discover hidden online anonymous marketplaces and messaging networks such as Silk Road (forced to cease operations in 2013), I2P, Agora, and Evolution, which have relatively few restrictions on the types of goods and services sellers can offer or that buyers can solicit. Bitcoin is the common currency used on these marketplaces due to its ability to preserve payment anonymity (Weimann 2016). Dark web applications also include those used by journalists, activists, and whistleblowing for file-sharing

such as SecureDrop, GlobalLeaks, and Wikileaks (Gadek et al. 2018). There are also social media networks on the dark web such as Galaxy2 that offer anonymous and censor-free alternatives to Twitter, Facebook, YouTube, etc. (Gehl 2016; Gadek et al. 2018). Additionally, there are real-time anonymous chat alternatives such as Tor Messenger, The Hub, OnionChat, and Ricochet.

Behavioral Aspects of Users

Users of the surface web tend to be law-abiding, well-intentioned individuals who are simply engaging in their routine daily tasks such as conducting basic internet searches or casually watching YouTube. The surface web is easy to navigate and does not present many technological challenges to most users. Surface web pages tend to be reliably available to users as long as they follow the conduct and use policies that are increasingly common, but they also tend to track their traffic, location, and IP address. Today, much of the surface web pages are embedded with advertising that relies on these tracking and identifying mechanisms. The majority of surface web users are aware of their acceptance of these conduct and use policies, and tacitly comply so that they can continue to use the sites. Surface web users are also increasingly becoming relatively immune to the advertising practices of many of these surface web pages, choosing to endure the ads so that they can continue on with the use of the site. Users of the deep web tend to be somewhat more technologically savvy in that deep web pages require users to first find and navigate the site, and also create accounts and maintain the use of usernames and passwords to enable them to continue to use specific sites. Additionally, some deep web sites such as academic databases require a user to be knowledgeable about common search techniques such as experimenting with different combinations of keywords, phrases, and date ranges. Users operate on the dark web in order to deliberately remain anonymous and untraceable. User behavior on the dark web is commonly associated with illegal activity such as cybercrime, drugs, weapons, money laundering, restricted

chemicals, hardcore pornography, contract killing, and coordination activity of terrorist groups, for example (Weimann 2016; Chertoff 2017). Many dark web users experience a sense of freedom and power knowing that they can operate online anonymously (Gehl 2016; Gadek et al. 2018).

Not all dark web users fall into this category, however; some users, for example, may wish to hide their identities and locations due to fear of violence and political retaliation, while still others simply use it because they believe that internet censorship is an infringement on their rights (Chertoff 2017). Nonetheless, dark web sites predominantly serve as underground marketplaces for the exchange of various illegal (or ethically questionable, such as hacker forums) products and services such as those mentioned above (Gadek et al. 2018).

Topology of the Web

There are two primary attempts at modeling the topology of the hyperlinks between the web pages of the surface web, the Jellyfish and Bow Tie models. The Jellyfish model (Siganos et al. 2006) depicts web pages as nodes that are connected to one another to varying degrees. Those that are strongly connected comprise the dense, main body of the jellyfish and there is a hyperlink path from any page within the core of the group to every other page within the core; whereas, the nodes that are loosely connected and do not have a clear path to every other page constitute the dangling tentacles. The Bow Tie model (Metaxas 2012) also identifies a strongly connected central core of web pages as nodes. This central core is the “knot” in the bow tie diagram, and two other large groups of web pages (the “bows”) reside on opposite sides of the knot. One of these “bows” consists of all of the web pages that link to the core, but do not have any pages from the core that link back to them. The other “bow” in the bow tie are the web pages that the core links to but do not link back out to the core. These groups are called the “in-” and “out-” groups, respectively, referring to the “origination”

and “termination” aspects of their hyperlinks; i.e., the “In-group” web pages originate outside of the strongly connected core, and link into it; while, the “Out-group” web pages are linked to from the core and terminate there. Similar to the Jellyfish model, the “In” and “Out” groups of the Bow Tie model also each have “tendrils,” which are the web pages that have links to and from the web pages within the larger group. The web pages within a tendril do not belong to the larger “In” or “Out” groups, but link to or from them for some reason, which means that within each tendril there exist both “origination” and “termination” links. There is a fourth group in the Bow Tie model, which is comprised of all of the web pages that are entirely disconnected from the bow tie, meaning that they are not linked in any way to or from the core. A final group of web pages within the Bow Tie model are called “Tubes,” which consists of web pages that are not part of the strongly connected core but link from the “In” bow to the “Out” bow; i.e., these web pages link the web pages within the “In” group to the web pages within the “Out” group.

Socio-technical Implications

Several socio-technical implications can be identified underlying the dark web. The interaction between users and dark web applications can create both good and bad situations. For example, the case of online anonymity that the dark web offers can help numerous groups of users, such as civilians, military personnel, journalists and their audiences, law enforcement, and whistleblowers and activists. But the same online anonymity can also help users to commit crimes and escape being held accountable for committing those crimes (Chertoff 2017). From a global perspective, policymakers need to continue to work together to understand the deep web and the dark web in order to develop better search methods aimed at rooting out this criminal activity while still maintaining a high level of privacy protection for noncriminal users. There are a few already existing legal frameworks in place, which these hacking and cybercrime tools violate, such as the

United States’ Computer Fraud and Abuse Act, which addresses interstate and international commerce and bars trafficking in unauthorized computer access and computer espionage. Additionally, the 2001 Convention on Cybercrime of the Council of Europe (known as the Budapest Convention) allows for international law enforcement cooperation on several issues. Yet there still exist wide variations in international approaches to crime and territorial limitations with regard to jurisdiction, and “international consensus remains elusive” with some countries such as Russia and China actively resisting the formation of international norms. At the same time, however, Russia, China, and Austria have passed some of their own strict laws concerning the dark web such as the forced collection of encryption keys from internet service providers and deanonymizing or blocking Tor and/or arresting users discovered to be hosting a Tor relay on their computer (Chertoff 2017). Furthermore, high levels of government censorship in some countries can actually push users to the dark web to find application alternatives.

Challenges

There are challenges to the surface web mainly in the forms of continually maintaining the development of standard search engines in order to optimize their effectiveness and deterring web spammers who utilize tricks to attempt to influence page ranking metrics. Another challenge to the surface web is that some people and organizations object to their sites or documents being included in the index and are lobbying for the government to come up with an easy way to maintain a right to be deindexed. The main challenge with regard to the deep web is the searchability of its contents. With vast amounts of documents being invisible to standard internet search engines, users are missing out on useful and perfectly legal information. Deep web search engines have been developed, but for the most part remain in the realm of academic and proprietary business use. There are several challenges surrounding the use of the dark web, such as the

lack of endpoint anonymity in peer-to-peer file-sharing, meaning that sometimes the host nodes and destination nodes can be identified and may therefore be subjected to legal action (Biddle et al. 2002). This can be most problematic if users use their work or university networks to create host nodes. Workplace and institutional policymakers have the challenge of devising, implementing, and maintaining pertinent safeguards. Policy and law on a global level is also a continual challenge for the dark web. Of particular concern is the challenge of identifying and preventing organizational attempts of terrorist activity on the dark web (Weimann 2016). However, finding and maintaining an appropriate balance between privacy and freedom of expression and crime prevention is a difficult endeavor, especially when it comes to reaching some type of international agreement on the issues (Gehl 2016). Another challenge is deciding on the appropriate level of governmental action with regard to the dark web. For example, some of the aspects of the dark web are considered by many to be beneficial, such as whistleblowing and positive types of hactivism (such as when hackers executed a coordinated effort to take down a child abuse website) (Chertoff 2017).

Future Directions

Various data and web mining techniques have been used for data collection and analysis to study the dark web. Some projects are ongoing such as The University of Arizona Dark Web project that focuses on terrorism. The project has been successful in generating an extensive archive of extremist websites, forums, and documents. The project recognizes the need, however, for continual methodology development due to the rapid reactive nature of the terrorist strategists (Weimann 2016). The United States Defense Advanced Research Projects Agency (DARPA) and the National Security Agency (NSA) also actively pursue projects for understanding the dark web and developing related applications (Weimann 2016). Researchers are also experimenting

with deep web social media analysis; developing tools to scan the network and analyze the text according to common techniques such as topic modeling, sentiment analysis, influence analysis, user clustering, and graph network analysis (Gadek et al. 2018).

Further Reading

- Biddle, P., England, P., Peinado, M., & Willman, B. (2002). The darknet and the future of content protection. In *Proceedings from ACM CCS-9 workshop, digital rights management* (pp. 155–176). Washington, DC: Springer.
- Chertoff, M. (2017). A public policy perspective of the dark web. *Journal of Cyber Policy*, 2(1), 26–38.
- Gadek, G., Brunessaux, S., & Pauchet, A. (2018). Applications of AI techniques to deep web social network analysis. [PDF file]. *North Atlantic Treaty Organization Science and Technology Organization*: Semantic Scholar. Retrieved from <https://pdfs.semanticscholar.org/e6ca/0a09e923da7de315b2c0b146cdf00703e8d4.pdf>.
- Gehl, R. W. (2016). Power/freedom on the dark web: A digital ethnography of the dark web social network. *New Media & Society*, 18(7), 1219–1235.
- Metaxas, P. (2012). Why is the shape of the web a bowtie? In *Proceedings from April 2012 International World Wide Web Conference*. Lyon: Wellesley College Digital Scholarship and Archive.
- Siganos, G., Tauro, S. L., & Faloutsos, M. (2006). Jellyfish: A conceptual model for the AS internet topology. *Journal of Communications and Networks*, 8(3), 339–350.
- Weimann, G. (2016). Going dark: Terrorism on the dark web. *Studies in Conflict & Terrorism*, 30(3), 195–206.

Sustainability

Andrea De Montis¹ and Sabrina Lai²

¹Department of Agricultural Sciences, University of Sassari, Sassari, Italy

²Department of Civil and Environmental Engineering and Architecture, University of Cagliari, Cagliari, Italy

Synonyms

Eco-development; Sustainable development

Definition/Introduction

Sustainability is a flexible and variously defined concept that – irrespective of the exact wording – encompasses the awareness that natural resources are finite, that social and economic development cannot be detached from the environment and that equity across space and time is required if development is to be carried on in the long term.

The concept, as well as its operative translations, was shaped across the years through several global conferences and meetings, in which state representatives agreed upon policy documents, plans, and goals. Hence, this institutional context must be kept in mind to understand properly the sustainability concept as well as the concerted efforts to integrate it within both regulatory evaluation processes, whereby the environmental effects of plans and projects are appraised, and voluntary schemes aiming at certifying the environmental “friendliness” of processes, products, and organizations.

From an operational standpoint, several attempts have been made at measuring sustainability through quantitative indicators and at finding aggregate, easily communicable, indices to measure progresses and trends. In this respect, the adoption of big data is key to the assessment of the achievement of sustainability goals by complex societies. Recently, the resilience concept has emerged in sustainability discourses; this is a soft and somewhat unstructured approach, which is currently gaining favor because deemed appropriate to deal with ever-evolving environmental and social conditions.

Origins of the Term: First Definitions and Interpretations

Although an early warning of the environmental impacts of unsustainable agriculture was already present in the 1962 book *Silent Spring* by Rachel Carson, the origins of the concept can be traced back to the beginning of the 1970s, when two first, notable, attempts at overcoming the long-standing view of the planet Earth as an unlimited source of

resources at the mankind’s disposal were made with both the book *The Limits to Growth*, which prompted the concept of carrying capacity of the planet, and the United Nations Conference on the Human Environment held in Stockholm, in which the tentative term “eco-development” was coined. Subsequently, an early definition, and possibly the first one, provided in 1980 by the International Union for Conservation of Nature (IUCN) in its World Conservation Strategy stated that sustainable development (SD) “must take account of social and ecological factors, as well as economic ones; of the living and non-living resource base; and of the long term as well as the short-term advantages and disadvantages of alternative actions.” Hence, sustainability cannot be detached from – to the contrary, it tends to identify itself with – sustainable development.

The most widely known and cited definition is, however, the one later provided by the Brundtland Commission in 1987, according to which SD is development that “meets the needs of the present without compromising the ability of future generations to meet their own needs,” often criticized in the ecologists’ and environmentalists’ circles because of its putting at the core the “needs” of human beings, both present and future generations. Notwithstanding, this broad definition of sustainability was often understood as synonymous with environmental sustainability, primarily concerned with the consumption of renewable resources within their regeneration capacity and of non-renewable resources at a slow rate so as not to prevent future generation from using them as well. Therefore, the two other pillars of sustainability (social and economic) implied in the definition by the IUCN were often left in the background.

Two significant and opposing standpoints about sustainability concern “strong” and “weak” sustainability. While strong sustainability assumes that natural capital cannot be replaced by man-made capital (comprising manufactured goods, technological advancement, and knowledge), weak sustainability assumes that natural capital can be substituted for man-made capital provided that the total stock of capital is maintained for future generations.

Sustainability and Institutions

The evolution of the concept of SD and its operative translations in the public domain are intertwined with the organization and follow-up of a number of mega-conferences, also known as world summits. A synopsis of these events, mostly organized by the United Nations, is reported in Table 1.

The first two conferences were held before – and heralded – the definition of SD, as they stressed the opportunity to limit human development when it affects negatively the environment. Rio +20 was the latest mega-conference on SD

and addressed the discussion of strategic SDGs, whose achievement should be properly encouraged and monitored.

Sustainability Measures, Big Data, and Assessment Frameworks

Agenda 21, Chapter 40, urged international governmental and nongovernmental bodies to conceive, design, and operationalize SD indicators and to harmonize them at the national, regional, and global levels. Similarly, the milestone document “The Future We Want” has recently

Sustainability, Table 1 Synopsis of the major conferences on sustainable development

Place, date, website	Name, acronym (short name)	Main products	Key issues
Stockholm, 5–16 June 1972, https://sustainabledevelopment.un.org/	United Nations Conference on the Human Environment, UNCHE	Declaration and action plan	Environmental consequences of human actions, environmental quality, improvement of the human environment for present and future generations, responsibility of the international community, safeguard of natural, especially renewable, resources
Nairobi, 10–18 May 1982	United Nations Environment Program, UNEP (Stockholm +10)	Declaration	Follow-up renovated recalls of the issues stressed in the UNCHE, focus on the unsatisfactory implementation of the UNCHE action plan
Rio de Janeiro, 3–14 June 1992, http://www.un.org/geninfo/bp/enviro.html	United Nations Conference on Environment and Development, UNCED (“Earth Summit”)	Rio declaration and action plan (Agenda 21)	Production of toxic substances, scarcity of water, and alternative energy sources; public transport systems; comprehensive action plan for the implementation of SD, monitoring role of the Commission of Sustainable Development (CSD)
New York, 23–28 June 1997	United Nations General Assembly Special Session, UNGASS (“Earth Summit II”)	Program for the further implementation of Agenda 21	Review of progress since the UNCED
Johannesburg, 26 August–6 September 2002, www.earthsummit2002.org/	World Summit on Sustainable Development, WSSD (Rio +10)	Declaration on SD (“political declaration”); civil society declaration	Sustainable development agreements in four specific areas: freshwater, sustainable energy, food security, and health
Rio de Janeiro, 20–22 June 2012, https://sustainabledevelopment.un.org/rio20	United Nations Conference on Sustainable Development, UNCSD (Rio +20)	“The future we want” report	Green economy and the institutional framework for SD; seven critical issues: jobs, energy, cities, food, water, oceans, and disasters; Sustainable Development Goals (SDGs)

emphasized “the importance of time-bound and specific targets and indicators when assessing progress toward the achievement of SDGs [...]”. According to Kwarto et al. (2016), several measures have been developed: over 500 indicators have been proposed by various governmental and nongovernmental organizations. Nearly 70 have been applied at the global level, over 100 at the national level, more than 70 at the subnational level, and about 300 at the local or metropolitan level. SD indicators should be designed as effective instruments able to support policy makers by communicating timely and precisely the performance of administrations at all levels. They should be relevant and describe phenomena communities need to know, easy to understand even by nonexperts, be reliable and correspond to factual and up-to-date situations, and be accessible and available when there is still time to react. Indicators often feed into broader frameworks, which are adopted to ascertain whether ongoing processes correctly lead to SD.

Following Rio +20, in 2015 the “2030 Agenda for Sustainable Development” was adopted through a resolution of the United Nations. The Agenda comprises a set of 17 Sustainable Development Goals (SDGs) together with 169 targets that have been envisaged to monitor the progress towards sustainability. Maarof (2015) argues that big data provide an unprecedented “opportunity to support the achievements of the SDGs”; the potential of big data to control, report, and monitor trends has been emphasized by various scholars, while Gijzen (2013) highlights three ways big data can help securing the sustainable future implied by the Agenda: first, they allow for modeling and testing different scenarios for sustainable conversion and enhancement of production processes; second, big data gathering, analysis, and modeling can help better understand current major environmental challenges such as climate change, or biodiversity loss; third, global coordination initiatives of big datasets developed by States or research centers (which also implies institutional coordination between big data collectors and analysts) would enable tracking each goal’s and target’s trend, hence the global progress

towards sustainability. Because of its openness and transparency, this continuous monitoring process enabled by big data would also entail improvements in accountability and in people empowerment (Maarof 2015). Big data analytics and context aware computing are expected to contribute to the project and development of innovative solutions integrating the Internet of Things (IoT) into smarter cities and territories (Bibri 2018). In this perspective, several attempts have been made to apply big data techniques to many other IoT domains, including healthcare, energy, transportation, building automation, agriculture, industry, and military (Ge et al. 2018). Some key issues need, however, to be considered, such as disparities in data availability between developed and developing countries (UN 2018), gender inequalities implying under- or over-representations, the need for public-private partnership in data production and collection.

Sustainability assessment is a prominent concept that groups any *ex-ante* process that aims to direct decision-making to sustainability. Sustainability assessment encompasses two dimensions: sustainability discourse and representation and decision-making context. Discourse is based on a pragmatic integration of development and environmental goals through constraints on human activities, which leads to representing sustainability in the form of disaggregated Triple Bottom Line (TBL) and composite variables. As for the second dimension, decisional contexts are disentangled in three areas: assessment (policies, plans, programs, projects, the whole Planet, and persistent issues), decision question (threshold, and choice), and responsible party (regulators, proponents, and third parties). In an institutional context, sustainability assessment constitutes the main theoretical and practical reference for several mandatory and voluntary processes of impact assessment. Mandatory processes consist of procedures regulated and imposed by specific laws and concerning the evaluation of the impacts over the environment caused by human activities. They include the North-American Environmental Impact Statement (EIS) and the European Environmental Impact Assessment (EIA) and Strategic

Environmental Assessment (SEA). EIS and EIA were introduced, respectively, in the USA in 1969 by the National Environment Policy Act and in Europe in 1985 by Directive 337/85/CE, whereas SEA was introduced in Europe in 2001 by Directive 2001/42/CE. EIS consists of a public procedure managed to clarify whether given plans or projects exert impacts over the environment and, if so, to propose proper mitigation strategies. As European counterparts, EIA and SEA have been introduced to assess and prevent environmental impacts generated by human activities connected respectively to certain projects of linear or isolated infrastructure or buildings and to the implementation of given plans and programs. Voluntary processes are spontaneous procedures originally carried out by private enterprises (and recently also by public bodies) to certify that products and processes comply with certain regulations regarding the quality of the Environmental Management System (EMS). Regulations include the Eco-Management and Audit Scheme (EMAS) elaborated by the European Commission and the 14,000 family of standards set by the International Standard Organization Technical Committee (ISO/TC 207). These processes imply relevant changes and continuous virtuous cycles and lead to the enhancement of: credibility, transparency, and reputation; environmental risk and opportunity management; environmental and financial performance; and employee empowerment and motivation. Techniques to appraise sustainability of products and processes include prominently Life Cycle Assessment (LCA). As standardized by ISO 14040 and 14044 norms, LCA is a methodology for measuring and disentangling environmental impacts associated with the life cycle of products, services, and processes. Historically, LCA was applied to evaluate the environmental friendliness of functional units, such as enterprises. Territorial LCA constitutes a new approach to the study of a territory, whereby the reference flow is the relationship between a territory and a studied land planning scenario. Territorial LCA enables to obtain two outputs: the environmental impacts and impacts associated with human activities in the territory.

Conclusions: Future of Sustainable Decision Making with Big Data

The operationalization of the concept of sustainability has so far heavily relied on attempts to identify proper indicators and measure them, possibly within the framework of dashboards (i.e., user-friendly software packages) and composite indices. Such indices are conceived to synthesize complex and multidimensional phenomena within an aggregate measure and include the Environmental Sustainability Index for the years 1999–2005 and the subsequent Environmental Performance Index, from 2006 onwards, both developed by Yale University and Columbia University, and the Human Development Index maintained by the United Nations. Other endeavors to quantify (un)sustainability of current development include communicative concepts aiming at raising awareness of the consequences of consumption choices and lifestyles, such as the Earth Overshoot Day, the Ecological Footprint, or the Carbon Footprint.

A different, and complementary approach to such quantitative tools to measure progresses towards, and divergence from, sustainability has taken place in the last years with the emergence of the resilience concept. As with “sustainability,” also with “resilience” different definitions coexist. In the engineering domain, resilience is grounded on the single equilibrium model and focuses on the pace at which a system returns to an equilibrium state after a disturbance. To the contrary, in the ecology domain, assumed that multiple equilibrium states can exist, resilience focuses on the amount of disturbance that a system can tolerate before shifting from a stability state to another, while reorganizing itself so as to maintain its functions in a changing environment. Similarly, when the resilience concept is applied to social-ecological systems, it carries the idea that such systems can endure disturbance by adapting themselves to a different environment through learning and self-organization, hence tending towards a new desirable state. Therefore, the resilience concept is often used in the context of mitigation and adaptation to climate change, and building

resilient communities and societies is increasingly becoming an imperative reference in sustainability discourses. These concepts call for the design of complex monitoring systems able to manage big data real time by pruning and visualizing immediately their trends and rationales. The achievement of SDGs for 2030 will constitute the political frontier of the extensive implementation of the IoT framework including sensors that capture real-time, continuous data, hardware devices for storing, software for processing big data and elaborating the analytics, and, ultimately, decision making tools able to select and, eventually, implement the necessary (re-)actions.

Cross-References

- [Environment](#)
- [Internet of Things \(IoT\)](#)
- [Social Sciences](#)
- [United Nations Educational, Scientific and Cultural Organization \(UNESCO\)](#)

Further Reading

- Bibri, S. E. (2018). The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. *Sustainable Cities and Society*, 38, 230–253.
- Ge, M., Bangui, H., & Buhnova, B. (2018). Big data for internet of things: A survey. *Future Generation Computer Systems*, 87, 601–614.
- Gijzen, H. (2013). Big data for a sustainable future. *Nature*, 502, 38.
- Kwatra, S., Kumar, A., Sharma, P., Sharma, S., Singhal, S. (2016). Benchmarking Sustainability Using Indicators: An Indian Case Study. *Ecological Indicators*, 61, 928–40.
- Maarof, A. (2015). *Big data and the 2030 agenda for sustainable development*. Draft report. https://www.unescap.org/sites/default/files/Final%20Draft_%20stock-taking%20report_For%20Comment_301115.pdf.
- UN. (2018). *Big data for sustainable development*. <http://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html>. Accessed 25 July 2018.

Sustainable Development

- [Sustainability](#)

Systemology

- [Systems Science](#)

Systems Science

Carolynne Hultquist
Geoinformatics and Earth Observation
Laboratory, Department of Geography and
Institute for CyberScience, The Pennsylvania
State University, University Park, PA, USA

Synonyms

[Systemology](#); [Systems theory](#)

Definition

Systems science is a broad interdisciplinary field that developed as an area of study in many disciplines that have a natural inclination to systems thinking. Instead of scientific reductionism which seeks to reduce things to their parts, systems thinking focuses on relating the parts as a holistic paradigm that considers interactions within the system and dynamic behavior. A system is not just a random collection of parts. Even Descartes, who argued for breaking complex problems into manageable parts, warned to be careful how one is breaking things apart. A system is connected by interactions and behaviors that provide meaning to the configuration that forms a system (Checkland 1981). Big data has volume and complexity that require processing to characterize and recognize patterns both within and between systems. Systems science has a long history of theoretical development in areas that deal with big data problems and is applied in a diversity of fields that employ systems thinking.

Introduction

Much of modern scientific thought focuses on reductionism which seeks to reduce things to

their parts. However, how parts fit together matters. Descartes even warned to be careful of how things are broken apart as a system is not just a random collection of parts. Aristotle argued that the whole is greater than the sum of its part and supported viewing a system as a whole (Cordon 2013). Systems thinking focuses on relating the parts in a holistic paradigm to consider interactions within the system and represent dynamic behavior from emerging properties. The systems science view focuses on the connection by interactions and behaviors that provide meaning to the configuration that forms a system (Checkland 1981). These theories could be applicable to characterizing complex big data of large scales over time with nonstandard sampling methods and data quality issues. The configuration of representation matters when attempting to recognize patterns, and big data analysis could benefit from a hierarchical systems modeling approach.

It is frequently asked in the data science field if meaningful general patterns of reality can be found in the data. Big datasets are often not a product of intentional sampling and are sometimes thought of as truly capturing the entire system or population. Some argue that big data is a well-defined system that enables new analysis by not relying on statistical sampling but looking at the “whole picture” as it is assumed to all be there due to the size of the dataset. However, regardless of how much data is available, data is simply a representation of reality, and we should not forget that portions of a system may inherently be excluded from collection. In addition, big data analysis presents problems of overfitting when too many parameters are set than is necessary and when meaningful variations of the system are erroneously specified as noise. This presents issues for future modeling as systematic trends may not be captured and reliability predicted. Systems science approaches can help to identify the system from the noise in order to improve modeling performance.

Systems science approaches could also encourage a critical awareness of the implications of data structure on analysis. Typically, data is fit into structured databases which is an imposition of a structure on the data which might cause data loss and not incorporate useful content on uncertainty. On the

other hand, denormalization is an approach that provides scalability and keeps data integrity by not reducing the data into smaller parts or making it necessary to do large-scale joins. There are yet many questions about the structure of datasets, and the form of analysis we utilize in data science fields as our understanding of patterns can be led astray in ill-defined systems. Theoretical assumptions on how we collect, represent, structure, and analyze the data should be critically considered in a systematic manner before determining what meaningful sense can be made of observed patterns.

Theoretical Frameworks

Systems theory is often used to refer to general systems theory which is the concept of applying systems thinking to all fields. Early research on systems science was driven by general systems research from the 1950s which endeavored to find a unified systems explanation of all scientific fields. It is an approach that was adapted over time to meet new application requirements, and the theories can be applied to understanding systems without the construction of a hypothesis-driven methodologically based analysis.

Systems science can be used to study systems from the simple to the complex. Complex systems or complexity science is a sister field that is specifically engaged with the philosophical and computational modeling challenges associated with complex behaviors and properties in systems (Mitchell 2009). As a framework, complexity builds off a holistic approach that argues it is impossible to predict emergent properties from only the initial parts as reduction does not give insight into the interactions of the system (Miller and Page 2007). Often, complex systems research prioritizes connections that can be represented using models such as agent-based models (ABM) or networks.

ABM and networks can grow in complexity by scaling up the model to more agents or adding in new ties which makes the processing more computationally intensive. In addition, there can be difficulties in parallelizing computation when a process cannot be split between multiple cores. Larger networks and more interactions can make

it difficult to identify specific patterns. Modeling agents or networks as a system, be it a social or physical model, can provide an environment to test the perceived patterns directly and perhaps allow for a comparison of the resulting model to raw data.

Systems modeling of big data could bring analysis beyond general correlation to causation. Standard big data analysis techniques lack explanatory power as they do not typically produce a hierarchical structure that leads to unification in order to make an argument from a general law. Instead, analysis focuses on causal models without an understanding of the system in which it occurred. Systems theory can provide a theoretical basis for creating systematic structures of modelled causal relationships that builds on other conditions and makes an argument for a generalized rules. In addition to building on complexity, chaos theory could find applications through big data analysis of chaotic systems that have so far only been systematically modeled. This theoretically branch of mathematics is applied as theory to many fields and is based on the concept that small changes of initial conditions in deterministic systems can have a significant impact on behavior in a dynamical system.

Key Applied Fields

The concepts of systems science are interdisciplinary and as a result it has been developed and applied in numerous fields. For example, Earth Systems Science is an interdisciplinary field that allows for a holistic consideration of the dynamic interactions of processes on Earth. Some applied fields, such as systems engineering, information systems, social systems, evolutionary economics, and network theory, gain significant attention by breakthroughs in these niche fields. The system and complexity sciences have long-standing traditions that could guide data scientists grappling with theoretical and applied problems.

Conclusion

Systems science transcends disciplinary boundaries and is argued to have unlimited scope (Warfield 2006). It can be considered a non-disciplinary paradigm for developing knowledge and insights in any discipline. It can also be viewed in relation to systems design in order to think critically about how parts seem to fit together. Essentially, systems science challenges reductionist thinking—whether theoretical or applied—by considering the dynamic interactions among elements in the system.

Systems science has a long history of developing theoretical frameworks. However, like big data, systems science has become a buzzword which has led those long in the field to question if new work in applied fields is grounded in theory. Basically, does the theory inform the practice? And if so, does the research outputs advance an understanding of either the applied system itself or systems thinking? The point is, systems science becoming a popular term does not necessarily advance the development of systems theories. Data scientists could provide benefit to the field by building off of the theoretical basis that informs the applied work.

Further Reading

- Checkland, P. (1981). *Systems thinking*. In *Systems practice*. Chichester, UK: Wiley.
- Cordon, C. P. (2013). System theories: An overview of various system theories and its application in healthcare. *American Journal of Systems Science*, 2(1), 13–22.
- Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An introduction to computational models of social life*. Princeton: Princeton University Press.
- Mitchell, M. (2009). *Complexity: A guided tour*. Oxford, UK: Oxford University Press. New York.
- Warfield, J. N. (2006). *An introduction to systems science*. Hackensack: World Scientific.

Systems Theory

- [Systems Science](#)

Tableau Software

Andreas Veglis
School of Journalism and Mass Communication,
Aristotle University of Thessaloniki,
Thessaloniki, Greece

Introduction

Tableau Software is a computer software company situated in Seattle (USA) that produces a series of interactive data visualization products (<http://tableau.com>). The company offers a variety of products that query relational databases, cubes, cloud databases, and spreadsheets and generate a variety of graph types. These graphs can be combined into dashboards and shared over the internet. The products utilize a database visualization language called VizQL (Visual Query Language), which is a combination of a structured query language for databases with a descriptive language for rendering graphics. VizQL is the core of the Polaris system, which is an interface for exploring large multidimensional databases. Special attention is given to the support of big data sets, since in recent years the demand of big data visualization has increased significantly. Tableau's users can work with big data without having advanced knowledge on query languages.

History

The company was founded by Chris Stolte, Christian Chabot, and Pat Hanrahan in Mountain View, California. The initial aim of the company was to commercialize research conducted by two of the founders at Stanford University's Department of Computer Science. The research included visualization techniques for exploring and analyzing relational databases and data cubes. Shortly after the company was moved to its present location, Seattle, Washington, it is worth noting that Tableau Software was one of the first companies to withdraw support from WikiLeaks after they started publishing US embassy cables.

Tableau and Big Data

Tableau supports more than 75 native data connectors and also a significant number of others via its extensibility options. Some examples of the supported connectors include SQL-based connections, NoSQL interfaces, open database connectivity (ODBC), and Web data connector. In order for its customers to have a fast interaction with their data, Tableau has developed a number of technologies, namely, hyper data engine, hybrid data architecture, and VizQL. Thus, Tableau offers a real-time interaction with the data.

Products

Tableau state-of-art data visualization is considered to be among the best of the business intelligence (BI) suites. BI can be defined as the transformation of raw data into meaningful and useful information for business analysis purposes. Tableau offers clean and elegant dashboards. The software utilizes drag-n-drop authoring in analysis and design processes. Tableau usually deals with at least two special dimensions, namely time and location. The use of special dimensions is quite interesting since some dimensions should be treated differently for more efficient analysis. For example, in maps and spatial analysis we usually employ location dimensions. Also in the case of time dimensions, they should be treated differently since in almost all cases information is relevant only in specific time context. Maps are considered to be one of the strongest features of Tableau products. Usually maps are quite difficult to develop by BI developers since most BI platforms do not offer strong support for maps. But this is not the case with Tableau since it incorporates excellent mapping functionalities. The latter is supported by regular updates of the offered maps and complimentary information (for example, income, population, and other statistical data) licensed from third parties. It is worth noting that Tableau can import, manipulate, and visualize data from various big data sources, thus following the current trend of working with big data.

As of the June 2020, Tableau Software offers seven main products: Tableau Desktop, Tableau Server, Tableau Online, Tableau Prep Builder, Tableau Mobile, Tableau Public, and Tableau Reader.

Tableau Desktop: It is a business intelligence tool that allows the user to easily visualize, analyze, and share large amounts of data. It supports importing data or connection with various types of database for automatic updates. While importing data the software also attempts to identify and categorize it. For instance, Tableau recognizes the country names and automatically adds information on the latitude and longitude of each country. This means that without any extra data entry

the user is able to use the program's mapping function to create a map of a specific parameter by country. The software also supports working with a number of different data sets. After the data has been imported by Tableau Desktop, the user can illustrate it with graphs, diagrams, and maps. The product is available for both Windows and MacOS platforms.

Tableau Server: Tableau Server is an enterprise-class business analytics platform that can scale up to hundreds of thousands of users. It supports distributed mobile, browser-based users, and Tableau Desktop clients to interact with Tableau workbooks published to the server from Tableau Desktop. The platform comprises four main components, namely the application server, the VizQL Server, the data server, and the backgrounder. The product can be installed on Windows and Linux servers, or it can be hosted in Tableau's data centers. Tableau Server is accessible through annual subscriptions.

Tableau Online: It is a hosted version running on the company's own multitenant cloud infrastructure of its data visualization product. Customers can upload data, create, maintain, and collaborate their visualization on the cloud (Tableau's servers). Tableau Online is usable from a Web browser, but is also compatible with Tableau Desktop. Tableau Online can connect to cloud-based data sources (for example, [Salesforce](#), Google BigQuery, and Amazon Redshift). Customers are also able to connect their own on-premise data.

Tableau Prep Builder: Introduced in 2018 this tool is used for the preparation of the data that will be analyzed with the help of another product of the Tableau ecosystem. It supports extraction, combination, and cleaning of data. As expected, Tableau Prep Builder works seamlessly with the other Tableau products.

Tableau Mobile: Tableau Mobile is a free mobile app (for iPhone, iPads, and Android phones and tablets) that allows users to access their data from mobile devices. Users can select, filter, and drill down their data and generally interact with the data using controls that are automatically optimized for devices with touch

screens. The Tableau Mobile can connect securely to Tableau Online and Tableau Server. It is worth noting that Tableau also offers the Tableau Vizable which is a free mobile app only available for iPads though which a user can access and explore.

Tableau Public: Tableau Public is a free available tool for any user who wants to create interactive data stories on the web. It is delivered as a service so it can be up and running immediately. Users are able to connect to data, create interactive data visualizations, and publish them directly to their website. Also they are able to guide readers through a narrative of data insights and allow them to interact with the data to make new discoveries. All visualizations are stored on the web and are visible by everyone. The product is available for both Windows and MacOS platforms.

Tableau Reader: Tableau Reader is a free desktop application that allows users to open, view, and interact with visualizations built in Tableau Desktop. It supports actions such as filter, drill down, and view details of the data as far as the author allows. Users are not able to edit or perform any interactions if the author has not built it. Tableau Reader is a simple way to share analytical insights.

Licenses

As of June 2020, Tableau offers annual subscriptions for individuals that include Tableau Desktop, Tableau Prep Builder, and one license for Tableau Server or Tableau Online. For teams and organizations there are similar subscriptions per user. Except the free products, all other products are available for downloading on a trial basis for 14 days. Also, Tableau Software offers free access to university students and instructors who want to utilize Tableau products in their courses.

It is worth mentioning that Tableau Software provides a variety of start-up guides (<http://www.tableau.com/support/>) training options to help customers get the most out of their data, and also

instructors who want to teach interactive visualizations with Tableau products.

Competitors

Today there are many companies that offer tools for creating interactive visualizations that can be considered competitors of Tableau Software. But its direct competitors are other BI platforms like Microsoft BI, SAP Business Objects, QlikView, IBM Cognos, Oracle Analytics Server, Sisense, Dundas BI, Microstrategy, Domo, and Birst.

Conclusions

Tableau's products are considered to be well designed and to be suited for nontechnical users. They are powerful, easy to use, highly visual, and aesthetically pleasant. By utilizing the free editions users can create complex interactive visualizations that can be employed for exploring complex data sets.

Cross-References

- [Business Intelligence](#)
- [Interactive Data Visualization](#)
- [Visualization](#)

Further Reading

- Lorth, A. (2019). *Visual analytics with Tableau*. Hoboken: Wiley.
- Milligan, J. (2019). *Learning Tableau 2019: Tools for business intelligence, data prep, and visual analytics* (3rd ed.). Birmingham: Packt Publishing.
- Sleeper, R. (2018). *Practical Tableau*. Newton: O'Reilly.
- Stirrup, J. (2016). *Tableau dashboard cookbook*. Birmingham: Packt Publishing.

Taxonomy

- [Ontologies](#)

Technological Singularity

Laurie A. Schintler and Connie L. McNeely
George Mason University, Fairfax, VA, USA

Technology is progressing at a rapid and unprecedented pace. Over the last half-century – and in the context of the “digital revolution” – global data and information storage, transmission, and processing capacities have ballooned, expanding super-exponentially rather than linearly. Moreover, profound advancements and breakthroughs in artificial intelligence (AI), genetics, nanotechnology, robotics, and other technologies and fields are being made continuously, in some cases on a day-to-day basis. Considering these trends and transformations, some scholars, analysts, and futurists envision the possibility of a *technological singularity* – i.e., a situation in which technological growth becomes unsustainable, resulting in a gradual or punctuated transition beyond even combined human capabilities alone (Vinge 1993). A technological singularity would be “an event or phase beyond which human civilization, and perhaps even human nature itself, would be radically changed” (Eden et al. 2012).

Theories about the technological singularity generally assume that accelerating scientific, industrial, social, and cultural change will give rise to a technology that is so smart that it can self-learn and self-improve. It would become ever-more intelligent with each iteration, possibly enabling an “intelligence explosion” and, ultimately, the rise of super-intelligence (Sandberg 2013). A “greater-than-intelligent” entity could be a single machine or a network of devices and humans with a collective intellect that far exceeds that of human beings (i.e., a social machine). Alternatively, it might arise from human intelligence amplification enabled by technologies such as computer/machine interfaces or whole brain emulation, or even profound advancements in biological science. At a technological singularity, there is the possibility of a “prediction horizon,” where projections become meaningless or impossible, where “we can no longer say anything

useful about the future” (Vinge 1993). Accordingly, regardless of how it manifests, a super-intelligent entity may be “the last invention that man need ever make” (Good 1966).

Big data has a big hand to play in the path to a technological singularity (Dresp-Langley et al. 2019). Indeed, the integration of big data, machine learning, and AI is increasingly described in terms of a data singularity (Arora 2018), arguably making for a disruptive intersection leading to the technological singularity. Therefore, new and expanding sources of structured and unstructured data have the potential to catalyze the transition to an intelligence explosion. For example, massive troves of streaming data produced by sensors, satellites, mobile devices, observatories, imaging, social media, and other sources help drive and shape AI innovations. The rapidly accelerating volume of big data also creates continual pressures to develop better and expanded computational and storage capabilities. Moreover, acceleration itself, referring to increasing rates of growth, is common across conceptions of the technological singularity (Eden et al. 2012). In recent years, related demands have led to an array of technological breakthroughs, e.g., quantum technology, amplifying in magnitudes of order the ability to acquire, process, and disseminate data and information. This situation has meant a better understanding and increased capabilities for modeling complex phenomena, such as climate change, dynamics in the cosmos, the physiology and pathology of diseases, and even the human brain’s inner-workings and human intelligence, which in turn fuel technological developments even further.

Alternatively, big data – or, more aptly, the “data tsunami” – can be viewed as an obstacle in the way to a technological singularity. The rate at which data and information are produced today is staggering, far exceeding management and analytical capacities – and the gap is widening. Thus, there is a deepening information overload. In this context, the share of imprecise, incorrect, and irrelevant data is amassing much faster than the proportion of data that can be used or trusted, contributing to increasing levels of “data smog” or “information pollution” (Shenk 1997). In other words, the signal-to-noise ratio is in a downward

spiral. Although new developments in artificial intelligence, such as deep learning, are significantly enhancing abilities to process and glean insights from big data, i.e., to see the signals, such technologies also are used in misinformation and disinformation, i.e., to produce noise. Consider, for example, “deepfake” images and videos or nefarious “bots” operating in social media.

Society has long been better at producing more data and information than it can consume. In fact, the information overload problem pre-dates the digital age by thousands of years (Blair 2010). Moreover, while new technologies and approaches always come on to the scene to help find, filter, organize, sort, index, integrate, appraise, and interpret data and information, there is always a predictable “knock-on” information explosion (de Solla Price 1961). For instance, the Web 2.0 era ushered in new tools and strategies for managing the information and data deluge produced in the first generation of the World Wide Web. This, in turn, resulted in an endless and ever-expanding collection of digital tags, ratings, annotations, and comments with each ensuing iteration of the Web.

For any information processing entity – whether an algorithm, a robot, an organization, a city, or the human brain – to be intelligent in the face of information overload, it must know “when to gather information...; where and in what form to store it; how to rework and condense it; how to index and give access to it; and when and on whose initiative to communicate it to others” (Simon 1996). That is, it must be a screener, compressor, synthesizer, and interpreter of information, with the capacity to listen and think, more than it speaks. In other words, the capacity to consume big data must exceed the capacity to produce it. Accordingly, the extent to which big data can facilitate a transition to a technological singularity hinges largely on the ability for machines (and humans) to manage the information overload.

- [Information Overload](#)
- [Information Quantity](#)
- [Machine Learning](#)

Further Reading

- Arora, A. (2018). Heading towards the data singularity. Towards Data Science. <https://towardsdatascience.com/heading-towards-the-data-singularity-829bfd82b3a0>
- Blair, A. M. (2010). *Too much to know: Managing scholarly information before the modern age*. New Haven/London: Yale University Press.
- de Solla Price, D. (1961). *Science since Babylon*. New Haven: Yale University Press.
- Dresp-Langley, B., Ekseth, O. K., Fesl, J., Gohshi, S., Kurz, M., & Sehring, H. W. (2019). Occam’s razor for big data? On detecting quality in large unstructured datasets. *Applied Sciences*, 9(15), 3065.
- Eden, A. H., Moor, J. H., Søraker, J. H., & Steinhart, E. (Eds.). (2012). *Singularity hypotheses: A scientific and philosophical assessment*. Heidelberg: Springer.
- Good, I. J. (1966). Speculations concerning the first ultra-intelligent machine. In *Advances in computers* (Vol. 6, pp. 31–88). Amsterdam: Elsevier.
- Sandberg, A. (2013). An overview of models of technological singularity. In *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future* (pp. 376–394). Hoboken: Wiley.
- Shenk, D. (1997). *Data smog*. New York: HarperCollins Publishers.
- Simon, H. A. (1996). Designing organizations for an information-rich world. *International Library of Critical Writings in Economics*, 70, 187–202.
- Vinge, V. (1993, March). *Technological singularity*. In VISION-21 symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute (pp. 30–31).

Telemedicine

Warren Bareiss

Department of Fine Arts and Communication Studies, University of South Carolina Upstate, Spartanburg, SC, USA

Cross-References

- [Artificial Intelligence](#)
- [Big Data Concept](#)

Overview

Telemedicine is the transmission and reception of health information and/or treatment from point to

point or among many points across space. It is used for diagnosis, treatment, and prevention as well as for research and continuing education. Other terms that are sometimes used synonymously, or at least with some semantic overlap, are “telehealth” and “e-health.” Telemedicine, in its various forms, brings healthcare services to remote locations, while making it possible to collect big data to better understand trends and service opportunities for poor and underserved populations far from urban medical centers.

The purpose of telemedicine is to increase accessibility to healthcare and health-related information where spatial distance is problematic leading to challenges not only pertaining to traveling distance but also time, expense, and shortage of medical professionals in areas that are not well served with regard to medical options or even accessibility to transportation. Telemedicine thus facilitates the flow of information among patients and providers for caregiving, on the one hand, and among healthcare professionals for training and coordination, on the other. Benefits include potentially faster diagnoses resulting in faster healing and more time available in treatment centers to care for patients who need to be admitted on-site.

Services provided via telemedicine are particularly needed in developing nations marked by high poverty, large rural communities, weak infrastructures including transportation systems, and chronic shortages of medical personnel – specialists in particular.

Telemedicine has three primary forms. Asynchronous telemedicine, called “search and forward” is the collection and storage of medical data for later retrieval. For example, a nurse practitioner at a remote location could take an X-ray, forward the image to a specialist via the Internet, and have the result in a matter of hours. Informational and reminder messages can also be programmed and send to patients.

The second form of telemedicine is the monitoring of patients in real time from a distant facility. Examples include use of step counters, gait sensors, and electromyography devices.

The third form is also synchronous wherein patients engage with healthcare providers “face to face” via a combination of audio and video technologies. Examples include patients being assisted at rural clinics and paramedics working with hospital emergency departments, transmitting vital signs and ascertaining if transportation to an emergency facility is required.

A wide range of telemedicine applications has been used globally including dermatology, psychotherapy, neonatal testing, genetic counseling, wound care, diabetes management, and neurology. Accuracy of synchronous telemedicine examinations has been shown to be comparable to that of traditional examinations when a bedside clinician is present to assist with the requisite technology.

Telemedicine in its current forms began in the early 1990s with the development of fiber optics necessary to carry large amounts of data to and from central hubs within respective systems. Because of its reliance upon visual data, radiology was the first major application.

Early in its development, telemedicine was still somewhat space biased, requiring patients and providers to communicate via dedicated facilities equipped with expensive videoconferencing technology. The cost and required technical support of such systems was prohibitive, especially in poor regions and in locations where reception was weak. Today, telemedicine systems are more commonly used due to the ease and relative low cost of using the Internet and readily available video cameras.

Regulation

Regulatory structures are closely associated with the expansion, acceptance, and use of telemedicine systems. Canada, for example, has a centralized system with universal accreditation across provinces. Regulation regarding reimbursement and malpractice are comparable with traditional care, in some cases providing higher reimbursement for telemedicine services as an incentive strategy.

The United States, on the other hand, lacks a centralized system so that licenses are not valid from state to state except in the case of the Veterans Administration. The issue reflects ongoing debate in the United States regarding the legitimacy of federal regulation in areas traditionally left to states, in this case, state medical boards.

In the United States, the telemedicine industry is represented by multiple lobbying agencies and professional organizations. One of the largest organizations advocating on behalf of telemedicine is the American Telemedicine Association (ATA) which provides information to professionals and to the general public, organizes an annual conference, and publishes a peer-reviewed journal: *Telemedicine and e-Health*. Among the most pressing issues for the ATA is promotion of interstate licensing agreements without which regional and national systems of telemedicine are currently impossible in the United States.

In 2015, the Center for Medicare and Medicaid Services (CMS) permitted Medicare to reimburse telemedicine providers, but maintained limitations so that only rural patients are eligible for telemedicine services. Furthermore, patients must receive treatment at approved “originating sites” such as health clinics, nursing facilities, and physicians’ offices. A final restriction is that Medicare reimbursement will only cover synchronous means of telemedicine between originating and distal sites.

The new ruling was welcomed, in part, by the ATA. The ATA’s reaction to the measure’s limitations, however, also reveals discord between the telemedicine industry and federal regulators. In response to the new ruling, for example, the ATA suggested that the CMS had slowed the proliferation of telemedicine when calling for more research instead of forging a multistate licensure agreement.

Like the ATA, the American Medical Association (AMA) also supports interstate licensure. In 2014, the AMA formally adopted the tripartite definition of telemedicine described above (synchronous delivery, remote monitoring, and store-and-forward systems) while joining in support of

the interstate licensing as proffered by the Interstate Licensing Board.

A fundamental barrier to multistate licensure is the fact that there is no single, universally agreed upon definition of what exactly telemedicine is or does. “Telemedicine” as a term is widely applied to include a plethora of technologies, applications, and personnel. cursory examination of medical literature on telemedicine reveals everything from a telephone conversation between doctor and patient to a host of specialized services emanating to a potentially vast public from a centralized source employing hundreds of healthcare providers.

Medicaid – administered on a state-by-state basis – has been much more supportive of telemedicine than has Medicare. According to a 2015 report published by the Center for Connected Health Policy (CCHP), 46 states currently provide some Medicaid reimbursement for synchronous, interactive telemedicine. Nine Medicaid programs support store-and-forward services apart from radiology, and 14 programs reimburse remote monitoring. Alaska, Minnesota, and Mississippi reimburse all three forms of telemedicine.

While the decentralized approach has slowed the growth of a national telemedicine system in the United States, more localized networks such as the South Carolina Telehealth Alliance and the Ohio Health Stroke Network support cooperative telemedicine endeavors among practitioners, medical centers, and government regulators within respective states.

Further, the so-called retail health movement in the United States has moved forward with telemedicine endeavors on a state-by-state basis, with California taking the lead. Pharmacy chain CVS began offering consultations with nurse practitioners via audio- and video-based technologies in selected California clinics in 2014. Other providers such as Kaiser Permanente, also based in California, are moving forward with retail-based telemedicine in conjunction with Target Corp. A similar partnership is under development between Sutter Express Care and Rite Aid in its California pharmacies.

Benefits

Beneficiaries of telemedicine include patients in rural locations where health practitioners are scarce. Transportation to medical facilities requires transportation, time, and money, any of which can be prohibitive to potential patients. Also, in nations where there are not enough doctors to serve the population regardless of the terrain, telemedicine can be used to more efficiently reach large numbers of patients. Finally, telemedicine is useful for illnesses where physical mobility itself is problematic.

From the start, the US military has been at the forefront of telemedicine, for example, among military personnel deployed to locations far from fully equipped medical facilities. Store-and-forward systems are used in multiple applications to collect images from cameras or cell phones. Images are then sent from an on-site medical provider via secure e-mail to a central manager who then distributes the information to a consultant and/or a specialist. The process is then reversed. Telemedicine used in this way has led to a reduction in unnecessary evacuations from remote locations which, in turn, reduce the need for replacement personnel. The speed in which military personnel receive consultation has been reduced from several weeks to a matter of hours. Also, the US Department of Veterans Affairs (VA) manages a robust telemedicine program stateside due to the high number of rural veterans, especially those who are losing mobility due to age and illness.

Telemedicine has demonstrated similar results among civilian populations by reducing unnecessary and costly transfer of patients from civilian populations in remote regions, again, speeding patient evaluations and promoting greater concentration on patients whose transfer requirements are more urgent and necessary. These factors also benefit insurers due to reduction of fees accruing from ambulance or helicopter transport and hospital readmission.

Numerous studies have reported that patient satisfaction with telemedicine is remarkably high. Patients appreciate convenient accessibility,

shorter stays in medical facilities, avoiding time away from work, and financial savings.

Barriers

Despite benefits to patients, providers, and insurers, development of telemedicine-based systems faces many impediments. Obstacles include lack of suitable infrastructure and national policies such as those described above, lack of expertise, perceived costs, and providers' unwillingness to adopt technologies and abandon old routines.

Lack of a generally agreed upon system of reimbursement is perhaps the greatest barrier to the proliferation of telemedicine systems worldwide, despite the fact that some research – albeit limited at this stage – has shown that over time telemedicine increases profit and costs less than traditional, in-person healthcare.

Another major concern is protection of privacy as information is shared and stored electronically. Besides ensuring that standard means of protecting patient privacy, such as HIPAA requirements, secure telemedicine systems require dedicated data storage and servers along with point-to-point encryption. Privacy is also a concern when delivery is not in a completely secure environment such as a kiosk at a retail center.

Further problematic issues include questions about the costs involving system maintenance and training. Also, the entire model of medicine at a distance via electronic technology puts much of the responsibility for health care in the hands of patients, for example in describing symptoms in isolation apart from a more integrated examination involving blood tests combined with the knowledge and trust built from traditional, long-term doctor-patient relationships.

Telemedicine and Big Data

Use of telemedicine technologies among large numbers of patients permits the collection of a vast amount of data across space and through time. Such data can be used to examine patterns and differences in health conditions within and

across geographical settings and to track changes across space and through time. Such data could be used to provide regional, national, and international profiles, while also pinpointing localized health issues, thus mitigating against the spread of health crises such as viral contagion. As such, big data gathered through telemedicine can be used to inform regional, national, and international healthcare policymaking in the areas of prediction, prevention, intervention, and promotion.

Conversely, big data can be used in conjunction with telemedicine platforms and applications to help determine treatment for individual patients and thereby improve medical service on a case-by-case basis. For example, big data can be used to predict emergency medical situations and the onset of specific diseases among individual patients when comparing aggregate data with patient records, family history, immediate symptoms, and so on. Further, use of easily portable telehealth devices makes such service relatively inexpensive when contrasted with long wait times and travel distances from remote locations to well-equipped healthcare facilities in population centers.

Looking Forward

Despite ongoing questions and regulatory issues, national and even global telemedicine technologies, practices, and uses seem poised for exponential growth due to the ubiquity of cell phones across the globe, even in many poor nations and regions severely lacking in healthcare providers. Furthermore, given their propensity to communicate via social media, today's adolescents are likely to welcome health care via telemedicine if messaging is easy to access and use, personalized, and is interactive. Easy access and affordability are not ends in themselves, however, as new questions will be raised particularly regarding appropriate training and oversight.

Although telemedicine does little to alleviate economic and social conditions that cause shortages of medical personnel in many parts of the world, it does appear to offer many

benefits to patients and providers in need of immediate, accessible, and cost-effective health care.

Cross-References

- [Data Sharing](#)
- [Data Storage](#)
- [Health Care Delivery](#)
- [Social Media](#)

Further Reading

- Achey, M., et al. (2014). Past, present, and future of telemedicine for Parkinson's disease. *Movement Disorders*, 29.
- American Telemedicine Association. <http://www.american-telemed.org>. Accessed May 2015.
- Azkari, A., et al. (2014). The 60 Most highly cited articles published in the *Journal of Telemedicine and Telecare* and *Telemedicine Journal and E-health*. *Journal of Telemedicine and Telecare*, 20, 871–883.
- Center for Connected Health Policy. <http://cchpca.org/telehealth-medicare-state-policy>. Accessed May 2015.
- Desjardins, D. (2015). Telemedicine comes to retail clinics. *Health Leaders Media*, 21/1, 1–3.
- Dougherty, J. P., et al. (2015). Telemedicine for adolescents with type 1 diabetes. *Western Journal of Nursing Research*, 36, 1199–1221.
- Hwang, J. S. (2014). Utilization of telemedicine in the U.S. military in a deployed setting. *Military Medicine*, 179, 1347–1353.
- Islam, R., et al. (2019). Portable health clinic: An advanced tele-healthcare system for unreached communities. In L. Ohno-Machado & B. Séroussi (Eds.), *MEDINFO 2019: Health and wellbeing e-networks for all*. International Medical Informatics Association (IMIA) and IOS Press, 616–619.
- Jelnes, R. (2014). Reflections on the use of telemedicine in wound care. *EWMA Journal*, 14, 48–51.
- Kalid, N., et al. (2018). Based real time remote health monitoring systems: A review on patients prioritization and related 'big data' using body sensors information and communication technology. *Journal of Medical Systems*, 42(2), 30.
- Leventhal, R. (2014). In Ohio, Optimizing Stroke Care with Telemedicine. Retrieved from <https://www.hcinnovationgroup.com/policy-value-based-care/article/13024392/in-ohio-optimizing-stroke-care-with-telemedicine>, Accessed 27 Nov 2020.
- Ma, L. V., et al. (2016). An efficient session weight load balancing and scheduling methodology for high-quality telehealth care service based on WebRTC. *The Journal of Supercomputing*, 72, 3909–3926.

- Sibson, L. (2014). The use of telemedicine technology to support pre-hospital patient care. *Journal of Parametric Practice*, 6, 344–353.
- Wenger, T. L. (2014). Telemedicine for genetic and neurologic evaluation in the neonatal intensive care unit. *Journal of Perinatology*, 34, 234–240.

Testing and Evaluation

- [Anomaly Detection](#)

The Big Data Research and Development Initiative (TBDRDI)

- [White House Big Data Initiative](#)

Time Series

- [Financial Data and Trend Prediction](#)

Time Series Analysis

- [Time Series Analytics](#)

Time Series Analytics

Erik Goepner
George Mason University, Arlington, VA, USA

Synonyms

[Time series analysis](#), [Time series data](#)

Introduction

Time series analytics utilize data observations recorded over time at certain intervals.

Subsequent values of time-ordered data often depend on previous observations. Time series analytics is, therefore, interested in techniques that can analyze this dependence (Box et al. 2015; Zois et al. 2015). Up until the second half of the twentieth century, social scientists largely ignored the possibility of dependence within time series data (Kirchgässner et al. 2012). Statisticians have since demonstrated that adjacent observations are frequently dependent in a time series and that previous observations can often be used to accurately predict future values (Box et al. 2015).

Time series data abound and are of importance to many. Physicists and geologists investigating climate change, for example, use annual temperature readings, economists study quarterly gross domestic product and monthly employment reports, and policy makers might be interested in before and after annual traffic accident data to determine the efficacy of safety legislation. Time series analytics can be used to forecast, determine the transfer function, assess the effects of unusual intervention events, analyze the relationships between variables of interest, and design control schemes (Box et al. 2015). Preferably, observations have been recorded at fixed time intervals. If the time intervals vary, interpolation can be used to fill in the gaps (Zois et al. 2015).

Of critical importance is whether the variables are stationary or nonstationary. Stationary variables are not time dependent (i.e., mean, variance, and covariance remain constant over time). However, time series data are quite often nonstationary. The trend of nonstationary variables can be deterministic (e.g., following a time trend), stochastic (i.e., random), or both. Addressing nonstationarity is a key requirement for those working with time series and is discussed further under “Challenges” (Box et al. 2015; Kirchgässner et al. 2012).

Time series are frequently comprised of four components. There is the trend over the long-term and, often, a cyclical component that is normally understood to be a year or more in length. Within the cycle, there can be a seasonal variation. And finally, there is the residual which includes all variation not explained by the trend, cycle, and seasonal components. Prior to the 1970s, only the

residual was thought to include random impact, with trend, cycle, and seasonal change understood to be deterministic. That has changed, and now it is assumed that all four components can be stochastically modeled (Kirchgässner et al. 2012).

The Evolution of Time Series Analytics

In the first half of the 1900s, fundamentally different approaches were pursued by different disciplines. Natural scientists, mathematicians, and statisticians generally modeled the past history of the variable of interest to forecast future values of the variable. Economists and other social scientists, however, emphasized theory-driven models with their accompanying explanatory variables. In 1970, Box and Jenkins published an influential textbook, followed in 1974 by a study from Granger and Newbold, that has substantially altered how social scientists interact with time series data (Kirchgässner et al. 2012).

The Box Jenkins approach, as it has been frequently called ever since, relies on extrapolation. Box Jenkins focuses on the past behavior of the variable of interest rather than a host of explanatory variables to predict future values. The variable of interest must be transformed so that it becomes stationary and its stochastic properties time invariant. At times, the terms *Box Jenkins approach* and *time series analysis* have been used interchangeably (Kennedy 2008).

Time Series Analytics and Big Data

Big Data has stimulated interest in efficient querying of time series data. Both time series and Big Data share similar characteristics relating to volume, velocity, variety, veracity, and volatility (Zois et al. 2015). The unprecedented volume of data can overwhelm computer memory and prevent processing in real time. Additionally, the speed at which new data arrives (e.g., from sensors) has also increased. The variety of data includes the medium from which it comes (e.g., audio and video) as well as differing sampling rates, which can prove problematic for data analysis. Missing data and incompatible sampling

rates are discussed further in the “Challenges” section below. Veracity includes issues relating to inaccurate, missing, or incomplete data. Before analysis, these issues should be addressed via duplicate elimination, interpolation, data fusion, or an influence model (Zois et al. 2015).

Contending with Massive Amounts of Data

Tremendous amounts of time series data exist, potentially overwhelming computer memory. In response, solutions are needed to lessen the effects on secondary memory access. Sliding windows and time series indexing may help. Both are commonly used; however, newer users may find the learning curve unhelpfully steep for time series indexing. Similarly, consideration should be given to selecting management schemes and query languages simple enough for common users (Zois et al. 2015).

Analysis and Forecasting

Time series are primarily used for analysis and forecasting (Zois et al. 2015). A variety of potential models exist, including autoregressive (AR), moving average (MA), mixed autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA). ARMA models are used with stationary processes and ARIMA models for nonstationary ones (Box et al. 2015). Forecasting options include regression and nonregression based models. Model development should follow an iterative approach, often executed in three steps: identification, estimation, and diagnostic checking. Diagnostic checks examine whether the model is properly fit, and the checks analyze the residuals to determine model adequacy. Generally, 100 or more observations are preferred. If fewer than 50 observations exist, development of the initial model will require a combination of experience and past data (Box et al. 2015; Kennedy 2008).

Autoregressive, Moving Average, and Mixed Autoregressive Moving Average Models

An autoregressive model predicts the value of the variable of interest based on its values from one or more previous time periods (i.e., its lagged value).

If, for instance, the model only relied on the value of the immediately preceding time period, then it would be a first-order autoregression. Similarly, if the model included the values for the prior two time periods, then it would be referred to as a second-order autoregression and so on. A moving average model also uses lagged values, but of the error term rather than the variable of interest (Kennedy 2008). If neither an autoregressive nor moving average process succeeds in breaking off the autocorrelation function, then a mixed autoregressive moving average approach may be preferred (Kirchgässner et al. 2012). AR, MA, and ARMA models are used with stationary time series, to include time series made stationary through differencing. However, the potential loss of vital information during differencing operations should be considered (Kirchgässner et al. 2012).

ARMA models produce unconditional forecasts, using only the past and current values of the variable. Because such forecasts frequently perform better than traditional econometric models, they are often preferred. However, blended approaches, which transform linear dynamic simultaneous equation systems into ARMA models or the inverse, are also available. These blended approaches can retain information provided by explanatory variables (Kirchgässner et al. 2012).

Autoregressive Integrated Moving Average (ARIMA) Models

In ARIMA models, also known as ARIMA (p, d, q), p indicates the number of lagged values of Y^* , which represents the variable of interest after it has been made stationary by differencing. d indicates the number of differencing operations required to transform Y into its stationary version, Y^* . The number of lagged values of the error term is represented by q . ARIMA models can forecast for univariate and multivariate time series (Kennedy 2008).

Vector Autoregressive (VAR) Models

VAR models blend the Box Jenkins approach with traditional econometric models. They can be quite helpful in forecasting. VAR models

express a single vector (of all the variables) as a linear function of the vector's lagged values combined with an error vector. The single vector is derived from the linear function of each variable's lagged values and the lagged values for each of the other variables. VAR models are used to investigate the potential causal relationship between different time series, yet they are controversial because they are atheoretical and include dubious assertions (e.g., orthogonal innovation of one variable is assumed to not affect the value of any other variable). Despite the controversy, many scholars and practitioners view VAR models as helpful, particularly VAR's role in analysis and forecasting (Kennedy 2008; Kirchgässner et al. 2012; Box et al. 2015).

Error Correction Models

These models attempt to harness positive features of both ARIMA and VAR models, accounting for the dynamic feature of time series data while also taking advantage of the contributions explanatory variables can make. Error correction models add theory-driven exogenous variables to a general form of the VAR model (Kennedy 2008).

Challenges

Nonstationarity

Nonstationarity can be caused by deterministic and stochastic trends (Kirchgässner et al. 2012). To transform nonstationary processes into stationary ones, the deterministic and/or stochastic trends must be eliminated. Measures to accomplish this include differencing operations and regression on a time trend. However, not all nonstationary processes can be transformed (Kirchgässner et al. 2012).

The Box Jenkins approach assumes that differencing operations will make nonstationary variables stationary. A number of unit root tests have been developed to test for nonstationarity, but their lack of power remains an issue. Additionally, differencing (as a means of eliminating unit roots and creating stationarity) comes with

the undesirable effect of eliminating any theory-driven information that might otherwise contribute to the model.

Granger and colleagues developed cointegrated procedures to address this challenge (Kirchgässner et al. 2012). When non-stationary variables are cointegrated, that is, the variables remain relatively close to each other as they wander over time, procedures other than differencing can be used. Examples of cointegrated variables include prices and wages and short- and long-term interest rates. Error correcting models may be an appropriate substitute for differencing operations (Kennedy 2008). Cointegration analysis has helped shrink the gap between traditional econometric methods and time series analytics, facilitating the inclusion of theory-driven explanatory variables into the modeling process (Kirchgässner et al. 2012).

Autocorrelation

Time series data are frequently autocorrelated and, therefore, violate the assumption of randomly distributed error terms. When autocorrelation is present, the current value of a variable serves as a good predictor of its next value. Autocorrelation can disrupt models such that the analysis incorrectly concludes the variable is statistically significant when, in fact, it is not (Berman and Wang 2012). Autocorrelation can be detected visually or with statistical techniques like the Durbin-Watson test. If present, autocorrelation can be corrected with differencing or by adding a trend variable, for instance (Berman and Wang 2012).

Missing Data and Incompatible Sampling Rates

Missing data occur for any number of reasons. Records may be lost, destroyed, or otherwise unavailable. At certain points, sampling rates may fail to follow the standard time measurement of the data series. Specialized algorithms may be necessary. Interpolation can be used as a technique to fill in missing data or to smooth the gaps between intervals (Zois et al. 2015).

Conclusion

Time series analytics utilizes data observations recorded over time at certain intervals, observations which often depend on each other. Time series analytics focuses on this dependence (Box et al. 2015; Zois et al. 2015). A variety of models exist for use in time series analysis (e.g., ARMA, ARIMA, VAR, and ECM). Of critical importance is whether the variables are stationary or non-stationary. Stationary variables are not time dependent (i.e., mean, variance, and covariance remain constant over time). However, time series data are quite often nonstationary. Addressing nonstationarity is a key requirement for users of time series (Box et al. 2015; Kirchgässner et al. 2012).

Cross-References

- [Core Curriculum Issues \(Big Data Research/Analysis\)](#)
- [Spatiotemporal Analytics](#)
- [Time Series Analytics](#)

Further Reading

- Berman, E., & Wang, X. (2012). *Essential statistics for public managers and policy analysts* (3rd ed.). Los Angeles: CQ Press.
- Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2015). *Time series analysis: Forecasting and control*. Hoboken: Wiley.
- Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Malden: Blackwell.
- Kirchgässner, G., Wolters, J., & Hassler, U. (2012). *Introduction to modern time series analysis* (2nd ed.). Heidelberg: Springer Science & Business Media.
- Zois, V., Chelmiss, C., & Prasanna, V. (2015). Querying of time series for big data analytics. In L. Yan (Ed.), *Handbook of research on innovative database query processing techniques* (pp. 364–391). Hershey: IGI Global.

Time Series Data

- [Time Series Analytics](#)

Transnational Crime

Louise Shelley

Terrorism, Transnational Crime, and Corruption
Center, George Mason University, Fairfax,
VA, USA

Transnational crime has expanded dramatically in the past two decades as criminals have benefited from the speed and anonymity of the cyber world and encrypted social media. Developments in technology have facilitated the growth of many forms of traditional crime as well as introduced cyber-dependent crime in which the crime is linked to pernicious items sold primarily on the dark web, such as ransomware, botnets, and trojans. These new tools deployed by criminals have permitted the theft of billions of private records, the theft of identities, and the enormous growth of illicit e-commerce. This criminal activity has expanded even more during the COVID-19 pandemic when individuals are isolated and spend greater amount of time on the Internet and on cell phones. The increase in this crime has required cyber security firms and law enforcement to rely more on large-scale data analytics to stop this crime and to locate and aid its victims and bring criminals to justice.

Transnational criminals have capitalized on the possibilities of the Internet, the deep and the dark web, and social media, especially its end-to-end encryption to expand their activities globally. Criminals are among the major beneficiaries of the anonymity of this new world of big data, providing them greater speed and outreach than previously. This phenomenal growth has occurred over the last two decades but has intensified particularly during the COVID-19 pandemic as individuals are more isolated and use their computers and cell phones more to engage with the outside world.

The use of big data for criminal uses can be divided into two distinct categories: cyber-enabled crime and cyber-dependent crime that can exist only in the cyber world. Cyber-enabled crimes include existing forms of crime that have

been transformed in scale or form by criminal use of the Internet, dark web, or social media. Included in this category are such crimes as drug trafficking, credit card fraud, human trafficking, and online sales of counterfeits, wildlife and antiquities. For example, dark websites have allowed bulk sales of narcotics facilitating impersonal interactions of drug traffickers and buyers. Silk Road, the first large online dark web drug marketplace, did billions of dollars in sales in its relatively short existence. Its replacement have continued to sell significant supplies of drugs online (Shelley 2018). During the COVID-19 pandemic, such cyber-enabled crimes as online fraud, dissemination of child abuse and pornography imagery, and sale of counterfeit medical products needed for the medical emergency have grown particularly.

Cyber-dependent crimes are defined as criminal activity in which a digital system is the target as well as the means of attack. Dark websites, accessed only by special software (e.g., Tor), sell these criminal tools such as ransomware, trojans, and botnets. Under this category of crime, information technology (IT) infrastructure can be disrupted, and data can be stolen on a massive scale using malware and phishing attacks. Many online cyber products are sold that can be used to extract ransoms, spread spam, and execute denial of service attacks. These same tools can lead to massive numbers of identity thefts and the theft of personal passwords facilitating intrusion into bank and other financial accounts and loss of large sums by victims. Ransomware sold online has been used to freeze the record systems of hospitals treating patients until ransom payments are made. Year-on-year growth is detected in cyber-dependent crimes and tens if not hundreds of millions of individuals were affected in 2020 through large-scale hacks and data breaches (Osborne 2020).

The availability of the Internet has provided for the dramatic expansion of customer access to purchase commercial sex and for exploiters to advertise victims of human trafficking. A major US government-funded computer research program, known as Memex, reported identified advertisement sales of approximately \$250 million spent

on posting more than 60 million advertisements for commercial sexual services in a 2-year period (Greenmeier 2015). The Memex tool that provides big data analytics for the deep web is now used to target the human trafficking criminals operating online. One human trafficking network, operating out of China, indicted by federal prosecutors was linked to hundreds of thousands of escort advertisements and 55 websites in more than 25 cities in the USA, Canada, and Australia. This case reveals how large-scale data analytics is now key to understanding the networks and the activities behind transnational organized crime (the USA et al. 2018).

Online and dark web sales as well as that conducted through social media are all facilitated by payment systems that process billions of transactions. The growth of global payments and the increased use of crypto currencies, many of them anonymized, make the identification of the account owners challenging. Therefore, finding the criminal transactions among the numerous international wire transfers, credit card, prepaid credit card, and crypto-currency transactions is difficult. Understanding the illicit activity requires the development of complex data analytics and artificial intelligence to ascertain the suspicious payments and link them with actual criminal activity.

Transnational criminals have been major beneficiaries of globalization and the rise of new technology. With their ability to use the Internet, deep and dark web, and social media to their advantage, capitalizing on anonymity and encryption, they have managed to advance their criminal objectives. Millions of individuals and institutions globally have suffered both personal and financial losses as law enforcement rarely possesses or keeps up with the advanced data analytics skills needed to counter the criminals' pernicious activities in social media and cyberspace.

Further Reading

Global Initiative Against Transnational Organized Crime. (2020). Crime and contagion: The impact of a pandemic on organized crime. <https://globalinitiative.net/wp-content/uploads/2020/03/CovidPB1rev.04.04.v1.pdf>. Accessed 22 Dec 2020.

Goodman, M. (2015). *Future crimes: Everything is connected, everyone is vulnerable and what we can do about it*. New York: Doubleday.

Greenmeier, L. (2015). Human traffickers caught on hidden Internet. <https://www.scientificamerican.com/article/human-traffickers-caught-on-hidden-internet/>. Accessed 22 Dec 2020.

Lusthaus, J. (2018). *Industry of anonymity: Inside the business of cybercrime*. Cambridge, MA: Harvard University Press.

Osborne, C. (2020). The biggest hacks, data breaches of 2020. <https://www.zdnet.com/article/the-biggest-hacks-data-breaches-of-2020/>. Accessed 22 Dec 2020.

Shelley, L. (2018). *Dark commerce: How a new illicit economy is threatening our future*. Princeton: Princeton University Press.

United States of America, Chen, Z. a.k.a. Chen, M., Zhou, W., Wang, Y. a.k.a. Sarah, Fu, T., & Wang, C.. (2018, November 15). <https://www.justice.gov/usao-or/press-release/file/1124296/download>. Accessed 22 Dec 2020.

Transparency

Anne L. Washington

George Mason University, Fairfax, VA, USA

Transparency is a policy mechanism that encourages organizations to disclose information to the public. Scholars of big data and transparency recognize the inherent power of information and share a common intellectual history. Government and corporate transparency, which is often implemented by releasing open data increases the amount of material available for big data projects. Furthermore, big data has its own need for transparency as data-driven algorithms support essential decisions in society with little disclosure about operations and procedures. Critics question whether information can be used as a control mechanism in an industry that functions as a distributed network.

Definition

Transparency is defined as a property of glass or any object that lets in light. As a governance mechanism, transparency discloses the inner

mechanisms of an organization. Organizations implement or are mandated to abide by transparency policies that encourage the release of information about how they operate. Hood and Heald (2006) use a directional typology to define transparency. Upward and downward transparency refers to disclosure within an organization. Supervisors observing subordinates is upward transparency, while subordinates observing the hierarchy above is downward transparency. Inward and outward transparency refers to disclosure beyond organizational boundaries. An organization aware of its environment is outward transparency, while citizen awareness of government activity is inward transparency. Transparency policies encourage the visibility of operating status and standard procedures.

First, transparency may compel information on operating status. When activities may impact others, organizations disclose what they are doing in frequent updates. For example, the US government required regular reports from stock exchanges and other financial markets after the stock market crash in 1929. Operating status information gives any external interest an ability to evaluate the current state of the organization.

Second, transparency efforts may distribute standard procedures in order to enforce ideal behaviors. This type of transparency holds people with the public trust accountable. For example, cities release open data with transportation schedules and actual arrival times. The planned information is compared to the actual information to evaluate behaviors and resource distribution. Procedural transparency assumes that organizations can and should operate predictably.

Disclosures allow comparison and review. Detailed activity disclosure of operations answers questions of who, what, when, and where. Conversely, disclosures can also answer questions about influential people or wasteful projects. Disclosure may emphasize predictive trends and retrospective measurement, while other disclosures may emphasize narrative interpretation and explanation.

Implementation

Transparency is implemented by disclosing timely information to meet specific needs. This

assumes that stakeholders will discover the disclosed information, comprehend its importance, and subsequently use it to change behavior. Organizations, including corporations and government, often implement transparency using technology which creates digital material used in big data.

Corporations release information about how their actions impact communities. The goal of corporate transparency is to improve services, share financial information, reduce harm to the public, or reduce reputation risks. The veracity of corporate disclosures has been debated by management science scholars (Bennis et al. 2008). On the one hand, mandatory corporate reporting fails if the information provided does not solve the target issue (Fung et al. 2007). On the other hand, organizations that are transparent to employees, management, stockholders, regulators, and the public may have a competitive advantage. In any case, there are real limits to what corporations can disclose and still remain both domestically and internationally competitive.

Governments release information as a form of accountability. From the creation of the postal code system to social security numbers, governments have inadvertently provided core categories for big data analytics (Washington 2014). Starting in the mid-twentieth century, legislatures around the world began to write freedom of information laws that supported the release of government materials on request. Subsequently, electronic government projects developed technology capabilities in public sector organizations.

Advances in computing have increased the use of big data techniques to automatically review transparency disclosures. Transparency can be implemented without technology, but often the two are intrinsically linked. One impact technology has on transparency is that information now comes in multiple forms. Disclosure before technology was the static production of documents and regularly scheduled reports that could be released on paper by request. Disclosure with technology is the dynamic streaming of real-time data available through machine-readable search and discovery. Transparency is

often implemented by releasing digital material as open data that can be reused with few limitations. Open data transparency initiatives disclose information in formats that can be used with big data methods.

Intellectual History

Transparency has its origins in economic and philosophical ideas about disclosing the activities of those in authority. In Europe, the intellectual history spans from Aristotle in fifth-century Greece to Immanuel Kant in eighteenth-century Prussia. Debates on big data can be positioned within these conversations about the dynamics of information and power. An underlying assumption of transparency is that there are hidden and visible power relationships in the exchange of information. Transparency is often an antidote to situations where information is used as power to control others.

Michel Foucault, the twentieth-century French philosopher, considered how rulers used statistics to control populations in his lecture on Governmentality. Foucault engaged with Jeremy Bentham's eighteenth-century descriptions of the ideal prison and the ideal government, both of which require full visibility. This philosophical position argues that complete surveillance will result in complete cooperation. While some research suggests that people will continue bad behavior under scrutiny, transparency is still seen as a method of enforcing good behavior.

Big data extends concerns about the balance of authority, power, and information. Those who collect, store, and aggregate big data have more control than those generating data. These conceptual foundations are useful in considering both the positive and negative aspects of big data.

Big Data Transparency

Big data transparency discloses the transfer and transformation of data across networks. Big data transparency brings visibility to the embedded power dynamic in predicting human behavior.

Analysis of digital material can be done without explicit acknowledgment or agreement. Furthermore, the industry that exchanges consumer data is easily obscured because transactions are all virtual. While a person may willingly agree to free services from a platform, it is not clear if users know who owns, sees, collects, or uses their data. The transparency of big data is described from three perspectives: sources, organizations, and the industry.

Transparency of sources discloses information about the digital material used in big data. Disclosure of sources explains which data generated on which platforms were used in which analysis. The flip side of this disclosure is that those who create user-generated content would be able to trace their digital footprint. User-generated content creators could detect and report errors and also be aware of their overall data profile. Academic big data research on social media was initially questioned because of opaque sources from private companies. Source disclosure increases confidence in data quality and reliability.

Transparency of platforms considers organizations that provide services that create user-generated content. Transparency within the organization allows for internal monitoring. While part of normal business operations, someone with command and control is able to view personally identifiable information about the activities of others. The car ride service Uber was fined in 2014 because employees used the internal customer tracking system inappropriately. Some view this as a form of corporate surveillance because it includes monitoring customers and employees.

Transparency of the analytics industry discloses how the big data market functions. Industry transparency of operations might establish technical standards or policies for all participating organizations. The World Wide Web Consortium's data provenance standard provides a technical solution by automatically tracing where data originated. Multi-stakeholder groups, such as those for Internet Governance, are a possible tool to establish self-governing policy solutions. The intent is to heighten awareness of the data supply chain from upstream content quality to downstream big data production. Industry transparency

of procedure might disclose algorithms and research designs that are used in data-driven decisions.

Big data transparency makes it possible to compare data-driven decisions to other methods. It faces particular challenges because its production process is distributed across a network of individuals and organizations. The process flows from an initial data capture to secondary uses and finally into large-scale analytic projects. Transparency is often associated with fighting potential corruption or attempts to gain unethical power. Given the influence of big data in many aspects of society, the same ideas apply to the transparency of big data.

Criticism

A frequent criticism of transparency is that its unintended consequences may thwart the anticipated goals. In some cases, the trend toward visibility is reversed as those under scrutiny stop creating findable traces and turn to informal mechanisms of communication.

It is important to note that a transparency label may be used to legitimize authority without any substantive information exchange. Large amounts of information released under the name of transparency may not, in practice, provide the intended result. Helen Margetts (1999) questions whether unfiltered data dumps obscure more than they reveal. Real-time transparency may lack meaningful engagement because it requires intermediary interpretation. This complaint has been lodged at open data transparency initiatives that did not release crucial information.

Implementation of big data transparency is constrained by complex technical and business issues. Algorithms and other technology are layered together, each with its own embedded assumptions. Business agreements about the exchange of data may be private, and release may impact market competition. Scholars question how to analyze and communicate what drives big data, given these complexities.

Other critics question whether what is learned through disclosure is looped back into the system for reform or learning. Information disclosed for transparency may not be channeled to the right places or people. Without any feedback mechanism, transparency can be a failure because it does not drive change. Ideally, either organizations improve performance or individuals make new consumer choices.

Summary

Transparency is a governance mechanism for disclosing activities and decisions that profoundly enhances confidence in big data. It builds on existing corporate and government transparency efforts to monitor the visibility of operations and procedures. Transparency scholarship builds on earlier research that examines the relationship between power and information. Transparency of big data evaluates the risks and opportunities of aggregating sources for large-scale analytics.

Cross-References

- [Business](#)
- [Data Governance](#)
- [Economics](#)
- [Privacy](#)
- [Standardization](#)

Further Reading

- Bennis, W. G., Goleman, D., & O'Toole, J. (2008). *Transparency: How leaders create a culture of candor*. San Francisco: Jossey-Bass.
- Fung, A., Graham, M., & Weil, D. (2007). *Full disclosure: The perils and promise of transparency*. New York: Cambridge University Press.
- Hood, C., & Heald, D. (Eds.). (2006). *Transparency: The key to better governance?* Oxford. New York: Oxford University Press.
- Margetts, H. (1999). *Information technology in government: Britain and America*. London: Routledge.
- Washington, A. L. (2014). Government information policy in the era of big data. *Review of Policy Research*, 31(4). <https://doi.org/10.1111/ropr.12081>.

Transportation Visualization

Xinyue Ye

Landscape Architecture & Urban Planning, Texas A&M University, College Station, TX, USA

The massive amounts of granular mobility data of people and transportation vehicles form a basic component in smart cities paradigm. The volume of available trajectory data has increased considerably because of the increasing sophistication and ubiquity of information and communication technology. The movement data records real-time trajectories sampled as a series of georeferenced points over transportation networks. There is an imperative need for effective and efficient methods to represent and examine the human and vehicle attributes as well as the contextual information of transportation phenomena in the comparative context. Visualizing the emerging large-scale transportation data can offer the stakeholders unprecedented capability to carry out data-driven urban system studies based on real-world flow information, in order to enhance communities in the twenty-first century. Nowadays, a large amount of transportation data sets are collected or distributed by administrations, companies, researchers, and volunteers. Some of them are available for public use, allowing to duplicate the results of a prior study using the same data and procedures. More datasets are not publicized due to the privacy, but the results of a prior study can be duplicated if the same procedures are followed but a similar transportation data is collected. The coming decades will witness more such datasets and especially openness of visual analytics procedures due to the increasing popularity of trajectory recording devices and citizen science. Open-source visual analytics represents a paradigm shift in transportation research that has facilitated collaboration across disciplines.

Since the early twenty-first century, the development of spatial data visualization for computational social science has been gaining momentum. As a multidimensional and multiscale phenomenon, transportation calls for scalable and

interactive visualization. To gain the insights from the heterogeneous and unstructured transportation data, the users need to conduct iterative, evolving information foraging, and sense making using their domain knowledge or collaborative thinking. Iterative visual exploration is fundamental in this process, which needs to be supported by efficient data management and visualization capabilities. Even simple tasks, such as smoothly displaying the heat maps of thousands of taxis' average, maximum, or minimum speed, cannot be easily completed with active user interactions without the appropriate algorithm design. Such operations require temporal and spatial data aggregations and visualizations with random access patterns, where advanced computational technologies should be employed. In general, an ideal visual analytics software system needs to offer the following: (1) powerful computing platform so that common users are not limited by their computational resources and can accomplish their tasks over daily-used computers or mobile devices; (2) easy-access gateway so that the transportation data can be retrieved, analyzed, and visualized by different user groups and their results can be shared and leveraged by others; (3) scalable data storage and management so that a variety of data queries can be immediately responded; (4) exploratory visualizations so that intuitive and efficient interactions can be facilitated; and (5) a multiuser system so that simultaneous operations are allowed by many users from different places.

Conventional transportation design software, such as TransCAD, Cube, and EMME, provides platforms for transportation forecasting, planning, and analysis with some visual representations of the results. However, these software packages are not specifically developed for big transportation data visualization. Domain practitioners, researchers, and decision-makers need to store, manage, query, and visualize big and dynamic transportation data. Therefore, transportation researchers demand handy and effective visual analytics software systems integrating scalable trajectory databases, intuitive and interactive visualization, and high-end computing resources. The availability of codes or publicly accessible

software will further play a transformative role for reproducible, replicable, and generalizable transportation sciences. Early transportation visualization with limited interaction capabilities relied on the abilities of traditional visualization methods, such as bar charts, line plots, and geographic information system mapping (e.g., heat maps or choropleths). However, many new methods and packages have been developed to visually explore trajectories data, using various visual metaphors and instant interactions, such as GeoTime, TripVista, FromDaDy, vessel movement, and TrajAnalytics. Such visualization will facilitate easy exploration of big trajectory data by an extensive community of stakeholders. Knowledge coproduction and community engagement can be strategically realized by using transportation visualization as a networking platform.

To conduct transportation visualization, two major technical components are needed: a robust database and an interactive visualization interface. A scalable database is a must for transportation data management, supporting fast computation over various data queries in a remote and distributed computing environment. An interactive visualization interface allows the researchers to query the data stored in database, to discover patterns, to generate hypotheses, and to share their insights with others. In addition to semantic zoom, brushing, and linking, the users can perform progressive visual exploration, for the purpose of interactively formulating hypotheses instead of testing hypotheses. Moreover, transportation visualization can reveal complex urban system phenomenon not identified otherwise. Public trajectory datasets can be made available on the cloud, so that researchers all over the world can stimulate transportation research and other activities to enhance the robustness and reliability of their urban and regional studies. On the other hand, the software can also be used privately by companies and research centers to manage their trajectory data on clouds or local clusters, where researchers with granted rights can access and visually analyze the data easily.

In summary, getting acquainted with data is the task of transportation visualization. In order to

develop a more spatially explicit transportation theory, it is first necessary to develop operational visualization that captures the spatial dynamics inherent in the datasets. The debates on transportation dynamics have been informed by, and to some extent inspired by, a parallel development of new visualization methods which has quantified the magnitude of human dynamics. The users can be released from the burden of computing capacity, allowing them to focus on their research questions on transport systems. Furthermore, transportation visualization can act as an outreach platform which can help government agencies to communicate the transportation planning and policies more effectively to the communities. With the coming age of digital twin, a digital replica of transportation system would also offer a new ecosystem for transportation visualization to play a critical role.

Cross-References

- [Cell Phone Data](#)
- [Visualization](#)

Further Reading

- Al-Dohuki, S., Kamw, F., Zhao, Y., Ye, X., Yang, J., & Jamonnak, S. (2019). An open source TrajAnalytics software for modeling, transformation and visualization of urban trajectory data. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 150–155). Piscataway: IEEE.
- Huang, X., Zhao, Y., Yang, J., Zhang, C., Ma, C., & Ye, X. (2016). TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), 160–169.
- Li, M., Ye, X., Zhang, S., Tang, X., & Shen, Z. (2017). A framework of comparative urban trajectory analysis. *Environment and Planning B*. <https://doi.org/10.1177/2399808317710023>.
- Pack, M. L. (2010). Visualization in transportation: Challenges and opportunities for everyone. *IEEE Computer Graphics and Applications*, 30(4), 90–96.
- Shaw, S., & Ye, X. (2019). Capturing spatiotemporal dynamics in computational modeling. In J. P. Wilson (Ed.), *The geographic information science & technology body of knowledge*. <https://doi.org/10.22224/gistbok/2019.1.6>.

Treatment

Qinghua Yang¹ and Yixin Chen²

¹Department of Communication Studies, Texas Christian University, Fort Worth, TX, USA

²Department of Communication Studies, Sam Houston State University, Huntsville, TX, USA

Treatment and Big Data

The Information Age has witnessed a rapid increase in biomedical information, which can lead to information overload and make information management difficult. One solution to managing large volume of data and reducing diagnostic errors is big data, which involve individuals' basic information, daily activities, and health conditions. Such information can come from different sources ranging from patients' health records, public health reports, and social media posts. After being digitalized, archived, and/or transformed, they grow into big data, which can serve as a valuable source for public health professionals and researchers to obtain new medical knowledge and find new treatment for diseases. For example, big data can be used to develop predictive models of clinical trials, and the results from such modeling and simulation can inform early decision-making for treatment.

One primary application of big data to medical treatment is personalized treatment, which refers to developing individualized therapies based on subgroups of patients who have a specific type of disease. Driven by the need for personalized medicine, big data have already been applied in the health care industry, particularly in treatment for cancer and rare diseases. Doctors can consult the databases to get advice on treatment strategies that might work for specific patients, based on the records of similar patients around the world. For instance, by interpreting biological data on childhood cancer patients, a research team at the University of Technology, Sydney, compared existing and previous patients' gene expressions and variations to assist clinicians at the bedside to determine the best treatment for patients. The

information revealed by huge medical databases (i.e., medical big data) can improve the understanding of potential risks and benefits of various treatments, accelerate the development of new medicines or treatments, and ultimately advance the health care quality.

There is a rapid development in academic research on personalized treatment using big data in recent years. For instance, a research project funded by the National Science Foundation used big data to explore better treatments for pain management, by creating a system that coordinates and optimizes all the available information including the pain data and taking into consideration a number of variables, such as daily living activities, marital status, and drug use. Similarly, in a project sponsored by the American Society of Clinical Oncology (ASCO), the researchers collected patients' age, gender, medications, and other illnesses, along with their diagnoses, treatment, and, eventually, date of death. Since the sheer volume of patients should overcome some data limitations (e.g., outliers), their overarching goal is to first accumulate data based on as many cancer patients as possible and then to analyze and quantify the data. Instead of giving a particular treatment for everyone for a particular disease, these projects aimed to personalize health care by having the treatment specifically for each patient.

Besides personalized treatment, there is also an increasing application of artificial intelligence and machine learning (ML) techniques in discovering new drugs and assessing their efficacy, as well as improving treatment. For instance, ranking methods, a new class of ML methods that can rank chemical structures based on their chances of clinical success, can be invaluable in prioritizing chemical compounds for screening and saving resources in developing compounds for new drugs. Although this new class of ML methods is promising during initial stages of screening, it turns out to be unsuitable for further development after several rounds of expensive (pre)clinical testing. Other important applications of ML techniques include self-organizing maps, multilayer perception, Bayesian neural networks, counter-propagation neural network, and support vector

machines, whose performance was found to have significant advantages compared to some traditional statistical methods (e.g., multiple linear regressions, partial least squares) in drug designs, especially in solving actual problems such as prediction of biological activities, construction of quantitative structure–activity relationships (QSAR) or quantitative structure–property relationships (QSPR) models, virtual screening, and the prediction of pharmacokinetic properties. Furthermore, IBM Watson’s cognitive computing helps physicians and researchers to analyze huge volume of data and focus on critical decision points, which is essential to improving the delivery of effective therapies and using the treatment to personalize therapies.

Controversy

Despite the promise of applying big data to medical treatment, some issues related to big data application are equally noteworthy. First, several questions remain unanswered regarding patients’ consent to have their data join the system, including how often the consent should be given, in what form the consent should be obtained, and whether it is possible to obtain true consent given the public’s limited knowledge about big data. Failure to appropriately answer these questions may engender ethical issues and misuse of big data in medical treatment.

Second, there is a gap between the curriculum in medical education and the need to integrate big data into better treatment decisions. Therefore, doctors may find information in medical big data, though overwhelming, not particularly relevant, for making treatment decisions. On the other hand, the algorithm models generated by big data analyses may not be transparent enough about why a specific treatment is recommended for certain patients, making these models like black boxes and thus may not be trusted by doctors.

Lastly, the translation process from academic research to medical practice is often expensive

and time-consuming. Thus, researchers in charge of advancing treatment are often constrained in their ability to improve people’s health and life quality. Also, some researchers can still publish their studies and get grants by following traditional methods, so there is not enough motivation for them to try innovative approaches such as big data for treatment development. The application of big data to treatment and health care can be held back, due to the long translation process of research to practice and the limited new knowledge generated by published studies following traditional methods.

Cross-References

- [Biomedical Data](#)
- [Health Care Delivery](#)
- [Health Informatics](#)
- [Patient Records](#)

Further Reading

- Agarwal, S., Dugar, D., & Sengupta, S. (2010). Ranking chemical structures for drug discovery: A new machine learning approach. *Journal of Chemical Information and Modeling*, 50(5), 716–731.
- DeGroff, C. G., Bhatikar, S., Hertzberg, J., Shandas, R., Valdes-Cruz, L., & Mahajan, R. L. (2001). Artificial neural network-based method of screening heart murmurs in children. *Circulation*, 103(22), 2711–2716.
- Duch, W., Swaminathan, K., & Meller, J. (2007). Artificial intelligence approaches for rational drug design and discovery. *Current Pharmaceutical Design*, 13(14), 1497–1508.
- Gertrudes, J. C., Maltarollo, V. G., Silva, R. A., Oliveira, P. R., Honorio, K. M., & Da Silva, A. B. F. (2012). Machine learning techniques and drug design. *Current Medicinal Chemistry*, 19(25), 4289–4297.
- Hoffman, S., & Podgurski, A. (2013). The use and misuse of biomedical data: Is bigger really better? *American Journal of Law & Medicine*, 39(4), 497–538.
- Liu, B. (2014). Utilizing big data to build personalized technology and system of diagnosis and treatment in traditional Chinese medicine. *Frontiers in Medicine*, 8(3), 272–278.
- Weingart, N. S., Wilson, R. M., Gibberd, R. W., & Harrison, B. (2000). Epidemiology of medical error. *BMJ*, 320(7237), 774–777.

U

United Nations Educational, Scientific and Cultural Organization (UNESCO)

Jennifer Ferreira
Centre for Business in Society, Coventry
University, Coventry, UK

United Nations Educational, Scientific and Cultural Organization (UNESCO), founded in 1945, is an agency of the United Nations (UN) which specializes in education, natural sciences, social and human sciences, culture, and communications and information. With 195 members, 9 associate members, and 50 field offices, working with over 300 international NGOs, UNESCO carries out activities in all of these areas, with the post-2015 development agenda underpinning their overall agenda.

As the only UN agency with a mandate to address all aspects of education, it proffers that education is at the heart of development, with a belief that education is fundamental to human, social, and economic development. It coordinates “Education for All” movement, a global commitment to provide quality basic education for all children, youth, and adults, monitoring trends in education and where possible make attempts to raise the profile of education on the global development agenda. For the natural sciences, UNESCO acts as an advocate for science as it focuses on encouraging international cooperation

in science as well as promoting dialogue between scientists and policy-makers. In doing so, it acts as a platform for dissemination of ideas in science and encourages efforts on crosscutting themes including disaster risk reduction, biodiversity, engineering, science education, climate change, and sustainable development. Within the social and human sciences, UNESCO plays a large role in promoting heritage as a source of identity and cohesion for communities. It actively contributes by developing cultural conventions that provide mechanisms for international cooperation. These international agreements are designs to safeguard natural and cultural heritage across the globe, for example, through designation as UNESCO World Heritage sites. The development of communication and sharing information is embedded in all their activities.

UNESCO has five key objectives: to attain quality education for all and lifelong learning; mobilize science knowledge and policy for sustainable development; address emerging social and ethical challenges; foster cultural diversity, intercultural dialogue, and culture of peace; and build inclusive knowledge societies through information and communication. Like other UN agencies, UNESCO has been involved in debates about the data revolution for development and the role that big data can play.

The data revolution for sustainable development is an international initiative designed to improve the quality of data and information that is generated and made available. It recognizes that

societies need to take advantage of new technologies and crowd-sourced data and improve digital connectivity in order to empower citizens with information that can contribute towards progress towards wider development goals. While there are many data sets available about the state of global education, it is argued that better data could be generated, even around basic measures such as the number of schools. In fact rather than focus on “big data” which has captured the attention of many leaders and policy-makers, instead more efforts should focus on “little data,” i.e., focus on data that is both useful and relevant to particular communities. Now discussions are shifting to identify which indicators and data should be prioritized.

UNESCO Institute for Statistics is the organization’s own statistic arm; however, much of the data collection and analysis that takes place here relies on much more conventional management and information systems which in turn relies on national statistical agencies which in many developing countries are often unreliable or heavily focused on administrative data (UNESCO 2012). This means that the data used by UNESCO is often out of date, or not detailed enough. While digital technologies have become widely used in many societies, more potential sources of data are generated (Pentland 2013). For example, mobile phones are now used as banking devices as well as for standard communications. Official statistics organizations are still behind in many countries and international organizations in that they have not developed ways to adapt and make use of this data alongside the standard administrative data already collected.

There are a number of innovative initiatives to make better use of survey data and mobile phone-based applications to collect data more efficiently and prove more timely feedback to schools, communities, and ministries on target areas such as enrolment, attendance, and learning achievement. UNESCO could make a significant contribution to a data revolution in education by investing in resources in collecting these innovations and making them more widely available to countries.

Access to big data for development, as with all big data sources, presents a number of ethical considerations based around the ownership of data and privacy. This is an area the UN recognizes that policy-makers will need to address to ensure that data will be used safely to address their objectives while still protecting the rights of people whom the data is about or generated from. Furthermore, there are a number of critiques of big data which make more widespread use of big data for UNESCO problematic: first that claims that big data are objective and accurate representations are misleading; not all data produced can be used comparably; there are important ethical considerations necessary about the use of big data; limited access to big data is exacerbating existing digital divides.

The Scientific Advisory Board of the Secretary-General of the United Nations which is hosted by UNESCO provided comments on the report on data revolution in sustainable development. It highlighted concerns over equity and access to data noting that the data revolution should lead to equity in access and use of data for all. Furthermore, it suggested that a number of global priorities should be included in any agenda related to the data revolutions: first that countries should seek to avoid contributing to a data divide between the rich and poor countries and secondly that there should be some form of harmonization and standardization of data platform to increase accessibility internationally, there should be national and regional capacity building efforts, and there should be a series of training institutes and training programs in order to develop skills and innovation in areas related to data generation and analysis (Manyika et al. 2011). A key point made here is that the quality and integrity of the data generated needs to be addressed, as it is recognized that big data often plays an important role in political and economic decision-making. Therefore a series of standards and methods for analysis and evaluation of data quality should be developed.

In the journal *Nature*, Hubert Gijzen, UNESCO Regional Science Bureau for Asia and

the Pacific, calls for more big data to help secure a sustainable future (Gijzen 2013). He argues that more data should be collected which can be used to model different scenarios for sustainable societies concerning a range of issues from energy consumption, improving water conditions, and poverty eradication. Big data, according to Gijzen, has the potential if coordinated globally between countries, regions, and relevant institutions to have a big impact on the way societies address some of these global challenges. The United Nations has begun to take actions to do this through the creation of the Global Pulse initiative bringing together experts from the government, academic, and private sectors to consider new ways to use big data to support development agendas. Global Pulse, a network of innovation labs which conduct research on Big Data for Development via collaborations between the governments, academic, and private sectors. The initiative is designed especially to make use of the digital data flood that has developed in order to address the development agendas that are at the heart of UNESCO, and the UN more broadly.

The UN Secretary-General's Independent Expert Advisory Group on the Data Revolution for Sustainable Development produced the report "A World That Counts" UN Secretary-General's Export Advisory Group on Data Revolution report in November 2014 suggested a number of key principles which should be sought regards to the use of data: data quality and integrity to ensure clear standards for use of data, data disaggregation to provide a basis for comparison, data timeliness to encourage a flow of high quality data for used in evidence-based policy-making, data transparency to encourage systems which allow data to make freely available, data usability to ensure data can be made user-friendly, data protection and privacy: to establish international and national policies and legal frameworks for regulating data generation and use, data governance and independence, data resources and capacity to ensure all countries have effective national statistical agencies, and finally data rights to ensure human rights remains

a core part of any legal or regulatory mechanisms that are developed with respect to big data (United Nations 2014). These principles are likely to influence UNESCOs engagement with big data in the future.

UNESCO, and the UN more broadly, acknowledge that technology has been, and will continue to be, a driver of the data revolution and a wider variety of data sources. For big data that is derived from this technology to have an impact, these data sources need to be leveraged in order to develop a greater understanding of the issues related to the development agenda.

Cross-References

- [International Development](#)
- [United Nations Educational, Scientific and Cultural Organization \(UNESCO\)](#)
- [World Bank](#)

Further Reading

- Gijzen, H. (2013). Development: Big data for a sustainable future. *Nature*, 52, 38.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. New York. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation. Accessed 12 Nov 14.
- Pentland, A. (2013). The data driven society. *Scientific American*, 309, 78–83.
- UNESCO (2012). Learning analytics. UNESCO Institute for Information Technologies Policy Brief. Available from <http://iite.unesco.org/pics/publications/en/files/3214711.pdf>. Accessed 11 Nov 14.
- United Nations (2014). A world that counts. United Nations. United Nations. Available from <http://www.unglobalpulse.org/IEAG-Data-Revolution-Report-A-World-That-Counts>. Accessed 28 Nov 14.

Unstructured Data

- [Data Integration](#)

Upturn

Katherine Fink
Department of Media, Communications, and
Visual Arts, Pace University, Pleasantville,
NY, USA

Introduction

Upturn is a think tank that focuses on the impact of big data on civil rights. Founded in 2011 as Robinson + Yu, the organization announced a name change in 2015 and expansion of its staff from two to five people. The firm's work addresses issues such as criminal justice, lending, voting, health, free expression, employment, and education. Upturn recommends policy changes with the aim of ensuring that institutions use technology in accordance with shared public values. The firm has published white papers, academic articles, and an online newsletter targeting policymakers and civil rights advocates.

Background

Principals of Upturn include experts in law, public policy, and software engineering. David Robinson was formerly the founding Associate Director of Princeton University's Center for Information Technology Policy, which conducts interdisciplinary research in computer science and public policy. Robinson holds a JD from Yale University's Law School and has reported for the *Wall Street Journal* and *The American*, an online magazine published by the American Enterprise Institute. Harlan Yu holds a PhD in Computer Science from Princeton University, where he developed software to make court records more accessible online. He has also advised the US Department of Labor on open government policies and analyzed privacy, advertising, and broadband access issues for Google. Aaron Rieke has a JD from the University of California Berkeley's Law School and has

worked for the Federal Trade Commission and the Center for Democracy and Technology on data security and privacy issues.

Cofounders Robinson and Yu began their collaboration at Princeton University as researchers on government transparency and civic engagement. They were among four coauthors of the 2009 *Yale Journal of Law & Technology* article "Government Data and the Invisible Hand," which argued that the government should prioritize opening access to more of its data rather than creating websites. The article suggested that "private parties in a vibrant marketplace of engineering ideas" were better suited to develop websites that could help the public access government data. In 2012, Robinson and Yu coauthored the *UCLA Law Review* article "The New Ambiguity of 'Open Government,'" in which they argued that making data more available to the public did not by itself make government more accountable. The article recommended separating the notion of open government from the technologies of open data in order to clarify the potential impacts of public policies on civic life.

Criminal Justice

Upturn has worked with the Leadership Conference, a coalition of civil rights and media justice organizations, to evaluate police department policies on the use of body-worn cameras. The organizations, noting increased interest in the use of such cameras following police-involved deaths in communities such as Ferguson (Missouri), New York City, and Baltimore, also cautioned that body-worn cameras could be used for surveillance, rather than protection, of vulnerable individuals. The organizations released a scorecard on body-worn camera policies of 25 police departments in November 2015. The scorecard included criteria such as whether body-worn camera policies were publicly available, whether footage was available to people who file misconduct complaints, and whether the policies limited the use of biometric technologies to identify people in recordings.

Lending

Upturn has warned of the use of big data by predatory lenders to target vulnerable consumers. In a 2015 report, “Led Astray,” Upturn explained how businesses used online lead generation to sell risky payday loans to desperate borrowers. In some cases, Upturn found that the companies violated laws against predatory lending. Upturn also found some lenders exposed their customers’ sensitive financial data to identity thieves. The report recommended that Google, Bing, and other online platforms tighten restrictions on payday loan ads. It also called on the lending industry to promote best practices for online lead generation and for greater oversight of the industry by the Federal Trade Commission and Consumer Financial Protection Bureau.

Robinson + Yu researched the effects of the use of big data in credit scoring in a guide for policymakers titled “Knowing the Score.” The guide endorsed the most widely used credit scoring methods, including FICO, while acknowledging concerns about disparities in scoring among racial groups. The guide concluded that the scoring methods themselves were not discriminatory, but that the disparities rather reflected other underlying societal inequalities. Still, the guide advocated some changes to credit scoring methods. One recommendation was to include “mainstream alternative data” such as utility bill payments in order to allow more people to build their credit files. The guide expressed reservations about “nontraditional” data sources, such as social network data and the rate at which users scroll through terms of service agreements. Robinson + Yu also called for more collaboration among financial advocates and the credit industry, since much of the data on credit scoring is proprietary. Finally, Robinson + Yu advocated that government regulators more actively investigate “marketing scores,” which are used by businesses to target services to particular customers based on their financial health. The guide suggested that marketing scores appeared to be “just outside the scope” of the Fair Credit Reporting Act, which requires agencies to notify consumers when their credit files have been used against them.

Voting

Robinson + Yu partnered with Rock the Vote in 2013 in an effort to simplify online voter registration processes. The firm wrote a report, “Connected OVR: a Simple, Durable Approach to Online Voter Registration.” At the time of the report, nearly 20 states had passed online voter registration laws. Robinson + Yu recommended that all states allow voters to check their registration statuses in real time. It also recommended that online registration systems offer alternatives to users who lack state identification, and that the systems be responsive to devices of various sizes and operating systems. Robinson + Yu also suggested that states streamline and better coordinate their online registration efforts. Robinson + Yu recommended that states develop a simple, standardized platform for accepting voter data and allow third-party vendors (such as Rock the Vote) to design interfaces that would accept voter registrations. Outside vendors, the report suggested, could use experimental approaches to reach new groups of voters while still adhering to government registration requirements.

Big Data and Civil Rights

In 2014, Robinson + Yu advised The Leadership Conference on “Civil Rights Principles for the Era of Big Data.” Signatories of the document included the American Civil Liberties Union, Free Press, and NAACP. The document offered guidelines for developing technologies with social justice in mind. The principles included an end to “high-tech profiling” of people through the use of surveillance and sophisticated data-gathering techniques, which the signatories argued could lead to discrimination. Other principles included fairness in algorithmic decision-making; the preservation of core legal principles such as the right to privacy and freedom of association; individual control of personal data; and protections from data inaccuracies.

The “Civil Rights Principles” were cited by the White House in its report, “Big Data: Seizing Opportunities, Preserving Values.” John Podesta,

Counselor to President Barack Obama, cautioned in his introduction to the report that big data had the potential “to eclipse longstanding civil rights protections in how personal information is used.” Following the White House report, Robinson + Yu elaborated upon four areas of concern in the white paper “Civil Rights, Big Data, and Our Algorithmic Future.” The paper included four chapters: Financial Inclusion, Jobs, Criminal Justice, and Government Data Collection and Use.

The Financial Inclusion chapter argued the era of big data could result in new barriers for low-income people. The automobile insurance company Progressive, for example, installed devices in customers’ vehicles that allowed for the tracking of high-risk behaviors. Such behaviors included nighttime driving. Robinson + Yu argued that many lower-income workers commuted during nighttime hours and thus might have to pay higher rates, even if they had clean driving records. The report also argued that marketers used big data to develop extensive profiles of consumers based on their incomes, buying habits, and English-language proficiency, and such profiling could lead to predatory marketing and lending practices. Consumers often are not aware of what data has been collected about them and how that data is being used, since such information is considered to be proprietary. Robinson + Yu also suggested that credit scoring methods can disadvantage low-income people who lack extensive credit histories.

The report found that big data could impair job prospects in several ways. Employers used the federal government’s E-Verify database, for example, to determine whether job applicants were eligible to work in the United States. The system could return errors if names had been entered into the database in different ways. Foreign-born workers and women have been disproportionately affected by such errors. Resolving errors can take weeks, and employers often lack the patience to wait. Other barriers to employment arise from the use of automated questionnaires some applicants must answer. Some employers use the questionnaires to assess which potential employees will likely stay in their jobs the longest. Some studies have suggested that longer commute times correlate to shorter-tenured

workers. Robinson + Yu questioned whether asking the commuting question was fair, particularly since it could lead to discrimination against applicants who lived in lower-income areas. Finally, Robinson + Yu raised concerns about “subliminal” effects on employers who conducted web searches for job applicants. A Harvard researcher, they noted, found that Google algorithms were more likely to show advertisements for arrest records in response to web searches of “black-identifying names” rather than “white-identifying names.”

Robinson + Yu found that big data had changed approaches to criminal justice. Municipalities used big data in “predictive policing,” or anti-crime efforts that targeted ex-convicts and victims of crimes as well as their personal networks. Robinson + Yu warned that these systems could lead to police making “guilt by association” mistakes, punishing people who had done nothing wrong. The report also called for greater transparency in law enforcement tactics that involved surveillance, such as the use of high-speed cameras that can capture images of vehicle license plates, and so-called stingray devices, which intercept phone calls by mimicking cell phone towers. Because of the secretive nature with which police departments procure and use these devices, the report contended that it was difficult to know whether they were being used appropriately. Robinson + Yu also noted that police departments were increasingly using body cameras and that early studies suggested the presence of the cameras could de-escalate tension during police interactions.

The Data Government and Use chapter suggested that big data tools developed in the interest of national security were also being used domestically. The DEA, for example, worked closely with AT&T to develop a secret database of phone records for domestic criminal investigations. To shield the database’s existence, agents avoided mentioning it by name in official documents. Robinson + Yu warned that an abundance of data and a lack of oversight could result in abuse, citing cases in which law enforcement workers used government data to stalk people they knew socially or romantically. The report also raised concerns about data collection by the US Census Bureau, which sought to lower the

cost of its decennial count by collecting data from government records. Robinson + Yu cautioned that the cost-cutting measure could result in undercounting some populations.

Newsletter

Equal Future, Upturn's online newsletter, began in 2013 with support from the Ford Foundation. The newsletter has highlighted news stories related to social justice and technology. For instance, *Equal Future* has covered privacy issues related to the FBI's Next Generation Identification system, a massive database of biometric and other personal data. Other stories have included a legal dispute in which a district attorney forced Facebook to grant access to the contents of nearly 400 user accounts. *Equal Future* also wrote about an "unusually comprehensive and well-considered" California law that limited how technology vendors could use educational data. The law was passed in response to parental concerns about sensitive data that could compromise their children's privacy or limit their future educational and professional prospects.

Cross-References

- [American Civil Liberties Union](#)
- [Biometrics](#)

- [e-commerce](#)
- [Financial Services](#)
- [Google](#)
- [Governance](#)
- [National Association for the Advancement of Colored People](#)
- [Online Advertising](#)

Further Reading

- Civil Rights Principles for the Era of Big Data. (2014, February). <http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html>.
- Robinson, D., & Yu, H. (2014, October). Knowing the score: New data, underwriting, and marketing in the consumer credit marketplace. https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf.
- Robinson + Yu. (2013). Connected OVR: A simple, durable approach to online voter registration. Rock the Vote. http://www.issuelab.org/resource/connected_ovr_a_simple_durable_approach_to_online_voter_registration.
- Robinson, D., Yu, H., Zeller, W. P., & Felten, E. W. (2008). Government data and the invisible hand. *Yale JL & Tech.*, 11, 159.
- The Leadership Conference on Civil and Human Rights & Upturn. (2015, November). Police body worn cameras: A policy scorecard. https://www.bwcscscorecard.org/static/pdfs/LCCHR_Upturn-BWC_Scorecard-v1.04.pdf.
- Upturn. (2014, September). Civil rights, big data, and our algorithmic future. <https://bigdata.fairness.io/>.
- Upturn. (2015, October). Led Astray: Online lead generation and payday loans. <https://www.teamupturn.com/reports/2015/led-astray>.
- Yu, H., & Robinson, D. G. (2012). The new ambiguity of 'open government'. *UCLA L. Rev. Disc.* 59, 178.

V

Verderer

- [Forestry](#)

Verification

- [Anomaly Detection](#)

Visible Web

- [Surface Web vs Deep Web vs Dark Web](#)

Visual Representation

- [Visualization](#)

Visualization

Xiaogang Ma
Department of Computer Science, University of
Idaho, Moscow, ID, USA

Synonyms

[Data visualization](#); [Information visualization](#);
[Visual representation](#)

Introduction

People use visualization for information communication. Data visualization is the study of creating visual representations of data, which bears two levels of meaning: the first is to make information visible and the second is to make it obvious for understand. Visualization is a pervasive existence in the data life cycle and recent trends is to promote the use of visualization in data analysis rather than use it only as a way to present the result. Community standards and open source libraries set the foundation for visualization of Big Data, and domain expertise and creative ideas are needed to put standards into innovative applications.

Visualization and Data Visualization

Visualization, in its literal meaning, is the procedure to form a mental picture of something that is not present to the sight (Cohen et al. [2002](#)). People can also illustrate such kind of mental pictures by using various visible media such as papers and computer screens. Seen as a way to facilitate information communication, the meaning of visualization can be understood at two levels. The first level is to make something to be visible and the second level is to make it obvious so it is easy to understand (Tufte [1983](#)). People's daily experience shows that graphics are easier to read and understand than words and numbers, such as the

use of maps in automotive navigation systems to show the location of an automobile and the road to the destination. The daily experience is approved by scientific discoveries. Studies on visual object perceptions explain such differentiation in reading graphics and texts/numbers: the human brain deciphers image elements simultaneously and decodes language in a linear and sequential manner, where the linear process takes more time than the simultaneous process.

Data are representations of facts and information is the meaning worked out from data. In the context of the Big Data, visualization is a crucial method to tackle the considerable needs of extracting information from data and presenting it. Data visualization is the study of creating visual representations of data. In practice, data visualization means to visually display one or more objects by combined use of words, numbers, symbols, points, lines, color, shading, coordinate systems, and more. While there are various choices of visual representations for a same piece of data, there are a few general guidelines that can be applied to establish effective and efficient data visualization. This first is to avoid distorting what the data have to say. That is, the visualization should not give a false or misleading account of the data. The second is to know the audience and serve a clear purpose. For instance, the visualization can be a description of the data, a tabulation of the records, or an exploration of the information that is of interest to the audience. The third is to make large datasets coherent. A few artistic designs will be required to present the data and information in an orderly and consistent way. The presidential, Senate, and House elections of the United States have been reported with well-presented data visualization, such as those on the website of *The New York Times*. The visualization on that website is underpinned by dynamic datasets and can show the latest records simultaneously.

Visualization in the Data Life Cycle

Visualization is crucial in the process from data to information. However, information retrieval is

just one of the many steps in the data life cycle, and visualization is useful through the whole data life cycle. In conventional understanding, a data life cycle begins with data collection and continues with cleansing, processing, archiving, and distribution. Those are from the perspective of data providers. Then, from the perspective of data users, the data life cycles continues with data discovery, access, analysis, and then repurposing. From repurposing, the life cycle may go back to the collection or processing step restarting the cycle. Recent studies show that there is another step called concept before the step of data collection. The concept step covers works such as conceptual models, logical models, and physical models for relational databases, and ontologies and vocabularies for Linked Data in the Semantic Web.

Visualization or more specifically data visualization provides support to different steps in the data life cycle. For example, the Unified Modeling Language (UML) provides a standard way to visualize the design of information systems, including the conceptual and logical models of databases. Typical relationships in UML include association, aggregation, and composition at the instance level, generalization and realization at the class level, and general relationships such as dependency and multiplicity. For ontologies and vocabularies in the Semantic Web, concept maps are widely used for organizing concepts in a subject domain and the interrelationships among those concepts. In this way a concept map is the visual representation of a knowledge base. The concept maps are more flexible than UML because they cover all the relationships defined in UML and allow people to create new relationships that apply to the domain under working (Ma et al. 2014). For example, there are concept maps for the ontology of the Global Change Information System led by the US Global Change Research Program. The concept maps are able to show that report is a subclass of publication, and there are several components in a report, such as chapter, table, figure, array, and image. Recent work in information technologies also enable online visualized tools to capture and explore concepts underlying collaborative science

activities, which greatly facilitate the collaboration between domain experts and computer scientists.

Visualization is also used to facilitate data archive, distribution, and discovery. For instance, the Tetherless World Constellation at Rensselaer Polytechnic Institute recently developed the International Open Government Dataset Catalog, which is a Web-based faceted browsing and search interface to help users find datasets of interest. A facet represents a part of the properties of a dataset, so faceted classification allows the assignment of the dataset to multiple taxonomies, and then datasets can be classified and ordered in different ways. On the user interface of a data center the faceted classification can be visualized as a number of small windows and options, which allows the data center to hide the complexity of data classification, archive and search on the server side.

Visual Analytics

The pervasive existence of visualization in the data life cycle shows that visualization can be applied broadly in data analytics. Yet, in actual practices visualization is often treated as a method to show the result of data analysis rather than as a way to enable the interactions between users and complex datasets. That is, the visualization as a result is separated from the datasets upon which the result is generated. Many of the data analysis and visualization tools scientists use in nowadays do not allow dynamic and live linking between visual representations and datasets, and when dataset changes, the visualization is no longer updated to reflect the changes. In the context of Big Data, many socioeconomic challenges and scientific problem facing the world are increasingly linked to the interdependent datasets from multiple fields of research, organizations, instruments, dimensions, and formats. Interactions are becoming an inherent characteristic of data analytics with the Big Data, which requires new methodologies and technologies of data visualization to be developed and deployed.

Visual analytics is a field of research to address the requests of interactive data analysis. It combines many existing techniques from data visualization with those from computational data analysis, such as those from statistics and data mining. Visual analytics is especially focused on the integration of interactive visual representations with the underlying computational process. For example, the IPython Notebook provides an online collaborative environment for interactive and visual data analysis and report drafting. IPython Notebook uses JavaScript Object Notation (JSON) as the scripting language, and each notebook is a JSON document that contains a sequential list of input/output cells. There are several types of cells to contain different contents, such as text, mathematics, plots, codes, and even rich media such as video and audio. Users can design a workflow of data analysis through the arrangement and update of cells in a notebook. A notebook can be shared with others as a normal file, or it can also be shared with the public using online services such as the IPython Notebook Viewer. A completed notebook can be converted into a number of standard output formats, such as HyperText Markup Language (HTML), HTML presentation slides, LaTeX, Portable Document Format (PDF), and more. The conversion is done through a few simple operations, so that means once a notebook is complete, a user only needs to press a few buttons to generate a scientific report. The notebook can be reused to analyze other datasets, and the cells inside it can also be reused in other notebooks.

Standards and Best Practices

Any applications of Big Data will face the challenges caused by the four dimensions of Big Data: volume, variety, velocity, and veracity. Commonly accepted standards or communities consensus are a proved way to reduce the heterogeneities between datasets under working. Various standards have already been used in application tackling scientific, social, and business issues, such as the aforementioned JSON for transmitting data with human-readable text, the

Scalable Vector Graphics (SVG) for two-dimensional vector graphics, and the GeoJSON for representing collections of georeferenced features. There are also organizations coordinating the works on community standards. The World Wide Web Consortium (W3C) coordinates the development of standards for the Web. For example, the SVG is an output of the W3C. Other W3C standards include the Resource Description Framework (RDF), the Web Ontology Language (OWL), and the Simple Knowledge Organization System (SKOS). Many of them are used for data in the Semantic Web. The Open Geospatial Consortium (OGC) coordinates the development of standards relevant to geospatial data. For example, the Keyhole Markup Language (KML) is developed for presenting geospatial features in Web-based maps and virtual globes such as Google Earth. The Network Common Data Form (netCDF) is developed for encoding array-oriented data. Most recently, the GeoSPARQL is developed for encoding and querying geospatial data in the Semantic Web.

Standards just enable the initial elements for data visualization, and domain expertise and novel ideas are needed to put standards into practice (Fox and Hendler 2011). For example, Google Motion Chart adapts the fresh idea of motion charts to extend the traditional static charts, and the aforementioned IPython Notebook allows the use of several programming languages and data formats through the use of cells. There are various programming libraries developed for data visualization, and many of them are made available on the Web. The D3.js is a typical example of such open source libraries (Murray 2013). The D3 here represents Data-Driven Documents. It is a JavaScript library using digital data to drive the creation and running of interactive graphics in Web browsers. D3.js based visualization uses JSON as the format of input data and SVG as the format for the output graphics. The OneGeology data portal provides a platform to browse geological map services across the world, using standards developed by both OGC and W3C, such as SKOS and Web Map Service (WMS). GeoSPARQL is a relatively newer standard for geospatial data but there are already

feature applications. The demo system of the Dutch Heritage and Location shows the linked open dataset of the National Cultural Heritage with more than 13 thousand archaeological monuments in the Netherlands. Besides the GeoSPARQL, GeoJSON and few other standards and libraries are also used in that demo system.

Cross-References

- [Data Visualization](#)
- [Data-Information-Knowledge-Action Model](#)
- [Interactive Data Visualization](#)
- [Pattern Recognition](#)

References

- Cohen, L., Lehericy, S., Chochon, F., Lemer, C., Rivaud, S., & Dehaene, S. (2002). Language-specific tuning of visual cortex? Functional properties of the visual word form area. *Brain*, 125(5), 1054–1069.
- Fox, P., & Hendler, J. (2011). Changing the equation on scientific data visualization. *Science*, 331(6018), 705–708.
- Ma, X., Fox, P., Rozell, E., West, P., & Zednik, S. (2014). Ontology dynamics in a data life cycle: Challenges and recommendations from a geoscience perspective. *Journal of Earth Science*, 25(2), 407–412.
- Murray, S. (2013). *Interactive data visualization for the web*. Sebastopol: O'Reilly.
- Tufte, E. (1983). *The visual display of quantitative information*. Cheshire: Graphics Press.

Vocabulary

- [Ontologies](#)

Voice Assistants

- [Voice User Interaction](#)

Voice Data

- [Voice User Interaction](#)

Voice User Interaction

Steven J. Gray
The Bartlett Centre for Advanced Spatial
Analysis, University College London,
London, UK

Synonyms

[Speech processing](#); [Speech recognition](#); [Voice assistants](#); [Voice data](#); [Voice user interfaces](#)

Introduction to Voice Interaction

Voice User Interaction is the study and methods of designing systems and workflows that process natural speech into commands and actions that can be carried out automatically on behalf of the user. The convergence of natural language processing research, machine learning and the availability of vast amounts of data, both written and voice, have allowed new ways of interaction into data discovery and traversing knowledge graphs. These interfaces allow users to control systems in a conversational way realizing the science fiction reality of conversing with computers to discover insights and control computing systems.

History of Conversational Interfaces

Interactive Voice Response systems (IVR) were developed, and first introduced commercially in 1973, which allowed users to interact with automated phone systems using Dual-tone multi-frequency signaling (DMTF) tones (Corkrey and Parkinson 2002). More commonly named “touch tone dialling” to select options allowing callers to navigate single actions through a tree on a phone call. As technology advanced and Computer Telephony Integration (CTI) was introduced into call centers, IVR systems became pseudo-interactive allowing the recognition of simple words to enable routing of calls to specific agents to handle the calls. Text to speech systems allow primitive

feedback to callers but are hardly intelligent as responses would be crafted based on the options selected on the call. Aural response statements would be pre-recorded or created from streams of words spoken by voice actors which often sound robotic or unnatural.

IVR Systems are regarded as the first wave of VUI but now with the advancement of natural language processing systems, machine learning models to detect, Named Entity Recognition, AI computational voice creation, and advanced speech recognition, it is now possible to create conversational interfaces that allow users to interact with larger Information Retrieval systems and Knowledge Graphs to answer requests. Vast amounts of voice data was collected through automated IVR systems such as Google Voice Local Search (“GOOG-411”), a service which allowed users to call Google to get search results over the phone (Bacchiani et al. 2008), allowing the training of AI systems that recognize speech patterns and accents to provide the various models needed to detect speech on device (Heerden et al. 2009).

The Introduction of Voice Assistants

Voice assistants are available on multiple surfaces, mobile devices, home hubs, watches, cars entertainment systems, and televisions for example allowing users to ask questions of these systems directly (Hoy 2018). Many devices can process the speech recording on a device, for example, for automated subtitles from video playback or transcription of messages instead of using touch interfaces. Voice detection is processed locally, on the device, to activate the assistant, using the cloud-based systems to send the users recording to cloud workflows to fulfill the request for data based on the query. Responses are created within seconds to create the illusion of conversation between the user and the device. Linking these voice assistants to knowledge bases allows answers to be digested by real time workflows and surface answers in conversational ways (Shalaby et al. 2020). The state of the conversation is saved in the cloud allowing users to ask follow up questions and the context being preserved to provide relevant

responses as would happen in natural conversation (Eriksson 2018).

Voice Assistants bring an additional form of modality into multimodal interfaces (Bourguet 2003). Voice will not necessarily replace other forms of interfaces but augment them and work in conjunction to provide the best interface for the user whatever the user's current circumstances are.

VUI design is diametrically opposed to Graphical User Interface design and standard User Experience paradigms. There are no graphical elements to select or present error states to the user, the system has to gracefully recover from error conditions by using prompts and responses. Simply, responding in a fashion of, "I'm sorry I don't understand that input" will confuse or frustrate a user, leading to users to interact less with such systems (Suhm et al. 2001). New design patterns for Voice Data systems have to adapt to the users inputs and lead them through the workflows to successful outcomes (Pearl 2016).

Expanding Voice Interfaces

Allowing developers to create customized third-party applications to surface on these virtual assistants unlocks the potential to expand these voice interfaces to systems that allow data entry and surface information to the user. In recent years, virtual assistants have also been expanded to screen calls on personal mobile phones on behalf of users as well as making bookings and organize events with local businesses. These businesses will soon replace adequate IVR systems to more conversational systems to interact with customers to free up employee time dealing with simple requests, bookings, and customer queries. In the near future, virtual assistants will be able to interface automatically with virtual business agents allowing machines to directly communicate with each other in natural language and becoming a digital personal assistant for the masses.

The combination of virtual assistants, knowledge graphs, and Voice User Interfaces has brought the science fiction dream of conversational computing to reality within the home. Being able to converse with computers and

information systems in a natural way will not only prove useful for many people in daily lives but also allow new surfaces for multimodal interaction giving users who find graphical user interfaces prohibitive or have various accessibility issues (Corbet and Weber 2016).

Further Reading

- Bacchiani, M., Beaufays, F., Schalkwyk, J., Schuster, M., & Strobe, B. (2008, March). Deploying GOOG-411: Early lessons in data, measurement, and testing. In *2008 IEEE international conference on acoustics, speech and signal processing* (pp. 5260–5263). IEEE.
- Bourguet, M. L. (2003). Designing and prototyping multimodal commands. In *Proceedings of human-computer interaction (INTERACT'03)* (pp. 717–720).
- Corbet, E., & Weber, A. (2016). What can I say? Addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services (MobileHCI'16)* (pp. 72–82). Association for Computing Machinery, New York. <https://doi.org/10.1145/2935334.2935386>.
- Corkrey, R., & Parkinson, L. (2002). Interactive voice response: Review of studies 1989–2000. *Behavior Research Methods, Instruments, & Computers*, 34, 342–353. <https://doi.org/10.3758/BF03195462>.
- Eriksson, F. (2018). Onboarding users to a voice user interface: Comparing different teaching methods for onboarding new users to intelligent personal assistants (Dissertation). Retrieved from <http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-149580>.
- Heerden, C. V., Schalkwyk, J., & Strobe, B. (2009). Language modeling for what-with-where on GOOG-411. In *Tenth annual conference of the international speech communication association*.
- Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1), 81–88.
- Pearl, C. (2016). *Designing voice user interfaces: Principles of conversational experiences*. Newton: O'Reilly.
- Shalaby, W., Arantes, A., GonzalezDiaz, T., & Gupta, C. (2020, June). Building chatbots from large scale domain-specific knowledge bases: Challenges and opportunities. In *2020 IEEE international conference on prognostics and health management (ICPHM)* (pp. 1–8). IEEE.
- Suhm, B., Myers, B., & Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction*, 8(1), 60–98.

Voice User Interfaces

► Voice User Interaction

Vulnerability

Laurie A. Schintler and Connie L. McNeely
George Mason University, Fairfax, VA, USA

Vulnerability is an essential and defining aspect of big data in today's increasingly digitalized society. Along with volume, variety, velocity, variability, veracity, and value, it is one of the "7 Vs" identified as principal determinant features in the conceptualizations of big data. Vulnerability is an integrated notion that concerns security and privacy challenges posed by the vast amounts, range of sources and formats, and the transfer and distribution of big data (Smartym Pro 2020). Data – whether a data feed, a trade secret, internet protocol, credit card numbers, flight information, email addresses, passwords, personal identities, transportation usage, employment, purchasing patterns, etc. – are accessible (Morgan 2015), and the nature of that accessibility can vary by purpose and outcome. Additionally, vulnerability encompasses the susceptibility of selected individuals, groups, and communities who are particularly vulnerable to data manipulation and inequitable application and use, and broader societal implications. To that end, vulnerability is a vital consideration regarding every piece of data collected (Marr 2016).

Vulnerability is a highly complex issue, with information theft and data breaches occurring regularly (DeAngelis 2018) and, more to the point, "a data breach with big data is a big breach" (Firican 2017). As an increasingly typical example, in the United States in September 2015, addresses and social security numbers of over 21 million current and former federal government employees were stolen, along with the fingerprints of 5.6 million (Morgan 2015). Data security and privacy challenges may occur through data leaks and cyber-attacks and the blatant hijacking and sale of data collected for legitimate purposes, for example, financial and medical data. Frankly, an unbreachable data repository simply does not exist (Morgan 2015).

Organizational data breaches can mean the theft of proprietary information, client data, and employee work and personal data. Indeed, banking and financial organizations, government agencies, and healthcare providers all face such big data security issues as a matter of course (Smartym Pro 2020).

The volume of data being collected about people, organizations, and places is exceptional, and what can be done with that data is growing in ways that would have been beyond imagination in previous years. "From bedrooms to boardrooms, from Wall Street to Main Street, the ground is shifting in ways that only the most cyber-savvy can anticipate" (Morgan 2015) and, in a world where sensitive data may be sold, traded, leaked, or stolen, questions of confidentiality and access, in addition to purpose and consequences, all underscore and point to problems of vulnerability.

Privacy and safety violations of personal data have been of particular concern. Data breaches and misuse have been broadly experienced by individuals and targeted groups in general, with broad social implications. Vulnerability addresses the fact that personal data – "the lifeblood of many commercial big data initiatives" – is being used to pry into individual behaviors and encourage purchasing behavior (Marr 2016). Personal data, including medical and financial data, are increasingly extracted and distributed via a range of connected devices in the internet-of-things (IoT). "As more sensors find their way into everything from smartphones to household appliances, cars, and entire cities, it is possible to gain unprecedented insight into the behaviors, motivations, actions, and plans of individuals and organizations" (Morgan 2015), such that privacy itself has less and less meaning over time. In other scenarios, vulnerability addresses the fact that personal data – "the lifeblood of many commercial big data initiatives" – is being used to pry into individual behaviors and encourage purchasing behavior (Marr 2016). In general, big data brings with it a range of security and privacy challenges – including the rampant sale of personal information on the dark web – and the proliferation of

data has left many people feeling exposed and vulnerable to the way their data is being violated and used (Experian 2017).

Traditionally disadvantaged and disenfranchised populations (e.g., the poor, migrants, minorities) are often the ones who are most vulnerable as a “result of the collection and aggregation of big data and the application of predictive analytics” (Madden et al. 2017). Indeed, there is an “asymmetric relationship between those who collect, store, and mine large quantities of data, and those whom data collection targets” (Andrejevic 2014). Moreover, algorithms and machine learning models have the propensity to produce unfair outcomes in situations where the underlying data used to develop and train them reflect societal gaps and disparities in the first place.

Big data and related analytics often are discussed as means for improving the world. However, the other side of the story is one of laying bare and generating vulnerabilities through information that can be used for nefarious purposes and to take advantage of various populations through, for example, identity theft and blanket marketing and purchasing manipulation. It is in this sense that an ongoing question is how to minimize big data vulnerability. That is, the challenge is to manage the vulnerabilities presented by big data now and in the future.

Cross-References

- [Big Data Concept](#)
- [Cybersecurity](#)
- [Ethics](#)
- [Privacy](#)

Further Reading

- Andrejevic, M. (2014). Big data, big questions| the big data divide. *International Journal of Communication*, 8, 17.
- DeAngelis, S. (2018). *The seven ‘Vs’ of big data*. <https://www.enterrasolutions.com/blog/the-seven-vs-of-big-data>.
- Experian. (2017). *A data powered future*. <https://www.experian.co.uk/blogs/latest-thinking/small-business/a-data-powered-future>.
- Firican, G. (2017). The 10 Vs of big data. *Upside*. <https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx>.
- Madden, M., Gilman, M., Levy, K., & Marwick, A. (2017). Privacy, poverty, and big data: A matrix of vulnerabilities for poor Americans. *Washington University Law Review*, 95, 53.
- Marr, B. (2016). Big data: The 6th ‘V’ everyone should know about. *Forbes*. <https://www.forbes.com/sites/bernardmarr/2016/12/20/big-data-the-6th-v-everyone-should-know-about/?sh=4182896e2170>.
- Morgan, L. (2015). 14 creepy ways to use big data. *InformationWeek*. <https://www.informationweek.com/big-data/big-data-analytics/14-creepy-ways-to-use-big-data/d/d-id/1322906>.
- Smartym Pro. (2020). *How to protect big data? The key big data security challenges*. <https://smartym.pro/blog/how-to-protect-big-data-the-main-big-data-security-challenges>.

Web Scraping

Bo Zhao

College of Earth, Ocean, and Atmospheric
Sciences, Oregon State University, Corvallis,
OR, USA

Web scraping, also known as web extraction or harvesting, is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis. Commonly, web data is scrapped utilizing Hypertext Transfer Protocol (HTTP) or through a web browser. This is accomplished either manually by a user or automatically by a bot or web crawler. Due to the fact that an enormous amount of heterogeneous data is constantly generated on the WWW, web scraping is widely acknowledged as an efficient and powerful technique for collecting big data (Mooney et al. 2015; Bar-Ilan 2001). To adapt to a variety of scenarios, current web scraping techniques have become customized from smaller *ad hoc*, human-aided procedures to the utilization of fully automated systems that are able to convert entire websites into well-organized data set. State-of-the-art web scraping tools are not only capable of parsing markup languages or JSON files but also integrating with computer visual analytics (Butler 2007) and natural language processing to simulate how human users browse web content (Yi et al. 2003).

The process of scraping data from the Internet can be divided into two sequential steps; acquiring web resources and then extracting desired information from the acquired data. Specifically, a web scraping program starts by composing a HTTP request to acquire resources from a targeted website. This request can be formatted in either a URL containing a GET query or a piece of HTTP message containing a POST query. Once the request is successfully received and processed by the targeted website, the requested resource will be retrieved from the website and then sent back to the give web scraping program. The resource can be in multiple formats, such as web pages that are built from HTML, data feeds in XML or JSON format, or multimedia data such as images, audio, or video files. After the web data is downloaded, the extraction process continues to parse, reformat, and organize the data in a structured way. There are two essential modules of a web scraping program – a module for composing an HTTP request, such as Urllib2 or selenium and another one for parsing and extracting information from raw HTML code, such as BeautifulSoup or Pyquery. Here, the Urllib2 module defines a set of functions to dealing with HTTP requests, such as authentication, redirections, cookies, and so on, while Selenium is a web browser wrapper that builds up a web browser, such as Google Chrome or Internet Explorer, and enables users to automate the process of browsing a website by programming. Regarding data extraction, BeautifulSoup is designed for

scraping HTML and other XML documents. It provides convenient Pythonic functions for navigating, searching, and modifying a parse tree; a toolkit for decomposing an HTML file and extracting desired information via `lxml` or `html5lib`. Beautiful Soup can automatically detect the encoding of the parsing under processing and convert it to a client-readable encode. Similarly, Pyquery provides a set of JQuery-like functions to parse xml documents. But unlike Beautiful Soup, Pyquery only supports `lxml` for fast XML processing.

Of the various types of web scraping programs, some are created to automatically recognize the data structure of a page, such as Nutch or Scrapy, or to provide a web-based graphic interface that eliminates the need for manually written web scraping code, such as Import.io. Nutch is a robust and scalable web crawler, written in Java. It enables fine-grained configuration, paralleling harvesting, robots.txt rule support, and machine learning. Scrapy, written in Python, is an reusable web crawling framework. It speeds up the process of building and scaling large crawling projects. In addition, it also provides a web-based shell to simulate the website browsing behaviors of a human user. To enable nonprogrammers to harvest web contents, the web-based crawler with a graphic interface is purposely designed to mitigate the complexity of using a web scraping program. Among them, Import.io is a typical crawler for extracting data from websites without writing any code. It allows users to identify and convert unstructured web pages into a structured format. Import.io's graphic interface for data identification allows user to train and learn what to extract. The extracted data is then stored in a dedicated cloud server, and can be exported in CSV, JSON, and XML format. A web-based crawler with a graphic interface can easily harvest and visualize real-time data stream based on SVG or WebGL engine but fall short in manipulating a large data set.

Web scraping can be used for a wide variety of scenarios, such as contact scraping, price change monitoring/comparison, product review collection, gathering of real estate listings, weather

data monitoring, website change detection, and web data integration. For examples, at a micro-scale, the price of a stock can be regularly scraped in order to visualize the price change over time (Case et al. 2005), and social media feeds can be collectively scraped to investigate public opinions and identify opinion leaders (Liu and Zhao 2016). At a macro-level, the metadata of nearly every website is constantly scraped to build up Internet search engines, such as Google Search or Bing Search (Snyder 2003).

Although web scraping is a powerful technique in collecting large data sets, it is controversial and may raise legal questions related to copyright (O'Reilly 2006), terms of service (ToS) (Fisher et al. 2010), and "trespass to chattels" (Hirschey 2014). A web scraper is free to copy a piece of data in figure or table form from a web page without any copyright infringement because it is difficult to prove a copyright over such data since only a specific arrangement or a particular selection of the data is legally protected. Regarding the ToS, although most web applications include some form of ToS agreement, their enforceability usually lies within a gray area. For instance, the owner of a web scraper that violates the ToS may argue that he or she never saw or officially agreed to the ToS. Moreover, if a web scraper sends data acquiring requests too frequently, this is functionally equivalent to a denial-of-service attack, in which the web scraper owner may be refused entry and may be liable for damages under the law of "trespass to chattels," because the owner of the web application has a property interest in the physical web server which hosts the application. An ethical web scraping tool will avoid this issue by maintaining a reasonable requesting frequency.

A web application may adopt one of the following measures to stop or interfere with a web scrapping tool that collects data from the given website. Those measures may identify whether an operation was conducted by a human being or a bot. Some of the major measures include the following: HTML "fingerprinting" that investigates the HTML headers to identify whether a visitor is malicious or safe (Acar et al. 2013); IP

reputation determination, where IP addresses with a recorded history of use in website assaults that will be treated with suspicion and are more likely to be heavily scrutinized (Sadan and Schwartz 2012); behavior analysis for revealing abnormal behavioral patterns, such as placing a suspiciously high rate of requests and adhering to anomalous browsing patterns; and progressive challenges that filter out bots with a set of tasks, such as cookie support, JavaScript execution, and CAPTCHA (Doran and Gokhale 2011).

Further Reading

- Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., & Preneel, B. (2013). Fpdetector: Dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC conference on computer & communications security*. New York: ACM.
- Bar-Ilan, J. (2001). Data collection methods on the web for infometric purposes – A review and analysis. *Scientometrics*, 50(1), 7–32.
- Butler, J. (2007). Visual web page analytics. Google Patents.
- Case, K. E., Quigley, J. M., & Shiller, R. J. (2005). Comparing wealth effects: The stock market versus the housing market. *The BE Journal of Macroeconomics*, 5(1), 1.
- Doran, D., & Gokhale, S. S. (2011). Web robot detection techniques: Overview and limitations. *Data Mining and Knowledge Discovery*, 22(1), 183–210.
- Fisher, D., McDonald, D. W., Brooks, A. L., & Churchill, E. F. (2010). Terms of service, ethics, and bias: Tapping the social web for CSCW research. *Computer Supported Cooperative Work (CSCW), Panel discussion*.
- Hirschey, J. K. (2014). Symbiotic relationships: Pragmatic acceptance of data scraping. *Berkeley Technology Law Journal*, 29, 897.
- Liu, J. C.-E., & Zhao, B. (2016). Who speaks for climate change in China? Evidence from Weibo. *Climatic Change*, 140(3), 413–422.
- Mooney, S. J., Westreich, D. J., & El-Sayed, A. M. (2015). Epidemiology in the era of big data. *Epidemiology*, 26(3), 390.
- O'Reilly, S. (2006). Nominative fair use and Internet aggregators: Copyright and trademark challenges posed by bots, web crawlers and screen-scraping technologies. *Loyola Consumer Law Review*, 19, 273.
- Sadan, Z., & Schwartz, D. G. (2012). Social network analysis for cluster-based IP spam reputation. *Information Management & Computer Security*, 20(4), 281–295.
- Snyder, R. (2003). Web search engine with graphic snapshots. Google Patents.
- Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques*. Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE. Melbourne, Florida, USA.

White House Big Data Initiative

Gordon Alley-Young

Department of Communications and Performing Arts, Kingsborough Community College, City University of New York, New York, NY, USA

Synonyms

[The Big Data Research and Development Initiative \(TBDRDI\)](#)

Introduction

On March 29, 2012, the White House introduced The Big Data Research and Development Initiative (TBDRDI) at a cost of \$200 million. Big data (BD) refers to the collection and interpretation of enormous datasets, using supercomputers running smart algorithms to rapidly uncover important features (e.g., interconnections, emerging trends, anomalies, etc.). The Obama Administration developed TBDRDI because having the large amounts of instantaneous data that is continually being produced by research and development (R&D) and emerging technology go unprocessed hurts the US economy and society. President Obama requested an all-hands-on-deck for TBDRDI including the public (i.e., government) and private (i.e., business) sectors to maximize economic growth, education, health, clean energy, and national security (Raul 2014; Savitz 2012). The administration stated that the private sector would lead by developing BD while the

government will promote R&D, facilitate private sector access to government data, and shape public policy. Several government agencies made the initial investment in this initiative to advance the tools/techniques required to analyze and capitalize on BD. TBDRDI has been compared by the Obama Administration to previous administrations' investments in science in technology that lead to innovations such as the Internet. Critics of the initiative argue that administration BD efforts need to be directed elsewhere.

History of the White House Big Data Initiative

TBDRDI is the White House's \$200 million federal agency funded initiative that seeks to secure the US's position as the world's most powerful and influential economy by channeling the information power of BD into social and economic development (Raul [2014](#)). BD is an all-inclusive name for the nonstop supply of sophisticated electronic data that is being produced by a variety of technologies and by scientific inquiry. In short, BD includes any digital file, tag or data that is created whenever we interact with technology, no matter how briefly (Carstensen [2012](#)). The dilemma posed by BD to the White House, as well as to other countries, organizations, and businesses worldwide, is that so much of it goes unanalyzed due to its sheer volume and the limits of our current technological tools to effectively store, organize, and analyze. Processing BD is not so simple because it requires supercomputing capabilities, some of which are still emerging. Experts argue that up until 2003, only 5-exabytes (EB) of data were produced; that number has since exploded to over five quintillion bytes of data (approximately 4.3 EB) every 2 days.

The White House Office of Science and Technology Policy (WHOSTP) announced TBDRDI in March 2012 in conjunction with the National Science Foundation (NSF), National Institutes of Health (NIH), US Geological Survey (USGS), and the Department of Defense (DoD) and Department of Energy (DoE). Key concerns to be addressed by TBDRDI are to manage BD by

significantly increasing the speed of scientific inquiry and discovery, bolstering national security, and overhauling US education. TBDRDI is the result of recommendations in 2011 by the President's Council of Advisors on Science and represents the US government's wish to get ahead of the wave BD and prevent a cultural lag by revamping its BD practices (Executive Office of the President [2014](#)). John Holdren, Director of WHOSTP, compared the \$200 million being invested in BD to prior federal investments in science and technology that are responsible for our current technological age (Scola [2013](#)). The innovations of the technology age ironically have created the BD that makes initiatives such as these necessary.

In addition to the US government agencies that helped to unveil TBDRDI, several other federal agencies had been requested to develop BD management strategies in the time leading up to and following this initiative. A US government fact sheet listed between 80 and 85 BD projects across a dozen federal agencies including, in addition to the departments previously mentioned, the Department of Homeland Security (DHS), Department of Health and Human Services (DHHS), and the Food and Drug Administration (FDA) (Henschen [2012](#)). The White House referred to TBDRDI as representing it placing its bet on BD meaning that the financial investment in this initiative is expected to yield a significant return for the country in coming years. To this end, President Obama has sought the involvement of public, private, and other (e.g., academia, non-governmental organizations) experts and organizations to work in a way that emphasizes collaboration. For spearheading TBDRDI and for choosing to stake the future of the country on BD, President Barack Obama has been dubbed the BD president by the media.

Projects of the White House Big Data Initiative

The projects included under the umbrella of TBDRDI are diverse, but they share common themes of emphasizing collaboration (i.e., to maximize resources and eliminate data overlap) and

making data openly accessible for its social and economic benefits. One project undertaken with the co-participation of NIH and Amazon, the world's largest online retailer, aims to provide public access to the 1,000 Genomes Project using cloud computing (Smith 2012). The 1,000 Genomes Project involved scientists and researchers sequencing the genomes of over 1,000 anonymous and ethnically diverse people between 2008 and 2012 in order to better treat illness and predict medical conditions that are genetically influenced. The NIH will deposit 200 terabytes (TB) of genomic data into Amazon's Web Services. According to the White House, this is currently the world's largest collection of human genetic data. In August 2014, the UK reported that it would undertake a 100,000 genomes project that is slated to finish in 2017. The NIH and NSF will cooperate to fund 15–20 research projects for a cost of \$25 million. Other collaborations include the DoE's and University of California's creation of a new facility as part of their Lawrence Berkeley National Laboratory called the Scalable Data Management, Analysis, and Visualization Institute (\$25 million) and the NSF and University of California, Berkeley's geosciences Earth Cube BD project (\$10 million).

The CyberInfrastructure for Billions of Electronic Records (CIBER) project is a co-initiative of the National Archives and Records Administration (NARA), the NSF, and the University of North Carolina Chapel Hill. The project will assemble decades of historical and digital-era documents on demographics and urban development/renewal. The project draws on citizen-led sourcing or citizen sourcing meaning that the project will build a participative archive fueled by engaged community members and not just by professional archivists and/or governmental experts. Elsewhere the NSF will partner with NASA's on its Global Earth Observation System of Systems (GEOSS), an international project to share and integrate Earth observation data. Similarly, the National Oceanic and Atmospheric Administration (NOAA) and NASA, who collectively oversee hundreds of thousands of environmental sensors producing reams of climate data, have partnered with Computer Science Corporation (CSC) to manage this climate data using their

ClimatEdge™ risk management suite of tools. CSC will collect and interpret the climate data and make it available to subscribers in the forms of monthly reports that anticipate how climate changes could affect global agriculture, global energy demand/production, sugar/soft commodities, grain/oilseeds, and energy/natural gas. These tools are promoted to help companies and consumers make better decisions. For example, fluctuating resource prices caused by climate changes will allow a consumer/business to find new supplies/suppliers in advance of natural disasters and weather patterns. Future goals include providing streaming data to advanced users of the service and expanding this service to other sectors including disease and health trends (Eddy 2014).

The DoD argues that it will spend \$250 million annually on BD. Several of its initiatives promote cybersecurity like its Cyber-Insider Threat program quick and precise targeting of cyber espionage threats to military computer networks. The DoD's cybersecurity projects also include developing cloud-computing capabilities that would retain function in the midst of an attack, programming languages that stay encrypted whenever in use, and security programs suitable for BD super-computer networks. In keeping with TBDRDI maxim to collaborate and share, the DoD has partnered with Lockheed Martin Corporation to provide the military and its partners with time-sensitive intelligence, surveillance, and reconnaissance data in what is being called a Distributed Common Ground System (DCGS). This project is touted as having the potential to save individual soldier's lives on the battlefield. Other defense-oriented initiatives under TBDRDI include how the Pentagon is working to increase its ability to extract information from texts over 100 times its current rates and Defense Advanced Research Projects Agency's (DARPA) development of XDATA (Raul 2014), a \$100 million program for sifting BD.

Influences of the Initiative and Expected Outcomes

The United Nations' (UN) Global Pulse Initiative (GPI) may have shaped TBDRDI (UN Global

Pulse 2012). Realizing in 2009–2010 that the data it relied upon to respond to global crises was outdated, the UN created its GPI to provide real-time data. In 2011, the proof of concept (i.e., primary project) phase began with the analysis of 2 years' worth of US and Irish social media data for mood scores/conversation indicators that could, in some cases, predict economic downturns 5 months out and economic upturns 2 months out. Success in this project justified opening GPI labs in Jakarta, Indonesia, and Kampala, Uganda.

Similarly in 2010, President Obama's Council of Advisors on Science and Technology urged focused investment in information technology (IT) to avoid overlapping efforts (Henschen 2012). This advice fit with 2010's existing cost-cutting efforts that were moving government work to less expensive Internet-based applications. TBDRDI, emerging from IT recommendations and after a period of economic downturn, differs from the so-called reality-based community (i.e., studying what has happened) of the Bush Administration by focusing instead on what will happen in the future. Some also argue that an inkling of TBDRDI can be seen as early as 2008 when then Senator Obama cosponsored a bipartisan online federal spending database bill (i.e., for USAspending.gov) and as a presidential candidate who actively utilized BD techniques (Scola 2013).

TBDRDI comes at a time when International Data Corporation (IDC) predicts that by 2020, over a third of digital information will generate value if analyzed. Making BD open and accessible will bring businesses an estimated three trillion dollars in profits. Mark Weber, President of US Public Sector for NetApp and government IT commentator, argues that the value of BD lies in transforming it into quality knowledge for increasing efficiency better informed decision-making (CIO Insight 2012). TBDRDI is also said to national security. Kaigham Gabriel, a Google executive and the next CEO and President of Draper Laboratory, argued that the cluttered nature of the BD field allows America's adversaries to hide and that field is becoming increasingly cluttered as it is estimated that government agencies generated one petabyte (PB) or one quadrillion bytes of data from 2012 to 2014 (CIO Insight 2012). One would need almost 14,552

64-gigabyte (GB) iPhones in order to store this amount of data. Experts argue that the full extent of technology/applications required to successfully manage the amounts BD that TBDRDI could produce now and in the future remains to be seen.

President Obama promised that TBDRDI would stimulate the economy and save taxpayer money, and there is evidence to indicate this. The employment outlook for individuals trained in mathematics, science, and technology is strong as the US government attempts to hire sufficient staff to carry out the work of TBDRDI. Hiring across governmental agencies requires the skilled work of deriving actionable knowledge from BD. This responsibility falls largely on a subset of highly trained professionals known as quantitative analysts or the quants for short. Currently these employees are in high demand and thus can be difficult to source as the US government must compete alongside private sector businesses for talent when the latter may be able to provide larger salaries and higher profile positions (e.g., Wall Street firms). Some have argued for the government to invest more money in the training of quantitative analysts to feed initiatives such as this (Tucker 2012).

In terms of cutting overspending, cloud computing (platform-as-a-service technologies) has been identified under TBDRDI as a means to consolidate roughly 1,200 unneeded federal data centers (Tucker 2012). The Obama Administration has stated that it will eliminate 40 % of federal data centers by 2015. This is estimated to generate a \$5 billion in savings. Some in the media applaud the effort and corresponding savings while some critics of the plan argue that the data centers be streamlined and upgraded instead. As of 2014, the US government reports that 750 data centers have been eliminated.

In January 2014, after classified information leaks by former NSA contractor Edward Snowden, President Obama asked the White House for a comprehensive review of BD that some argue dampened the enthusiasm for TBDRDI (Raul 2014). The US does not have a specific BD privacy law leading critics to claim a policy deficit. Others point to the Federal Trade Commission (FTC) Act, Section 5 that prohibits unfair or deceptive acts or practices in or affecting commerce as being firm enough to handle any untoward business practices that might

emerge from BD while flexible enough to not hinder the economy (Raul 2014). Advocates note that the European Union (EU) has adopted a highly detailed privacy policy that has done little to foster commercial innovation and economic growth (Raul 2014).

Conclusion

Other criticism argues that TBDRDI, and the Obama Administration by default, actually serves big business instead of individual consumers and citizens. In support of this argument, critics argue that the administration pressured communications companies to provide more affordable and higher speeds of mobile broadband. As of the summer of 2014, Hong Kong has the world's fastest mobile broadband speeds that are also some of the most affordable with South Korea second and Japan third; the US and its neighbor Canada are not even in the top ten list of fastest mobile broadband speed countries. Supporters of the administration cite that the Obama Administration has instead chosen to emphasize its unprecedented open data initiatives under TBDRDI. The US Open Data Action Plan emphasizes making high-priority US government data both mobile and publically accessible while Japan is reported to have fallen behind in open-sourcing its BD, specifically in providing access to their massive stores of state/local data, costing its economy trillions of yen.

Cross-References

- Big Data
- Cloud Computing
- Cyberinfrastructure (U.S.)
- National Oceanic and Atmospheric Administration

References

- Carstensen, J. (2012). *Berkeley group digs in to challenge of making sense of all that data*. Retrieved from http://www.nytimes.com/2012/04/08/us/berkeley-group-tries-to-make-sense-of-big-data.html?_r=0.
- CIO Insight (2012). *Can government IT meet the big data challenge?* Retrieved from <http://www.cioinsight.com/>

[c/a/Latest-News/Big-Data-Still-a-Big-Challenge-for-Government-IT-651653/](http://www.whitehouse.gov/the-press-office/2012/05/01/big-data-still-a-big-challenge-for-government-it-651653/).

- Eddy, N. (2014). *Big data proves alluring to federal IT pros*. Retrieved from <http://www.eweek.com/enterprise-apps/big-data-proves-alluring-to-federal-it-pros.html>.
- Executive Office of the President (2014). *Big data: Seizing opportunities, preserving values*. Retrieved from https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.
- Henschen, D. (2012). *Big data initiative or big government boondoggle?* Retrieved from <http://www.informationweek.com/software/information-management/big-data-initiative-or-big-government-boondoggle/d/d-id/1103666?>.
- Raul, A.C. (2014). *Don't throw the big data out with the bath water*. Retrieved from http://www.politico.com/magazine/story/2014/04/dont-throw-the-big-data-out-with-the-bath-water-106168_full.html?print#.U_PA-lb4bFI.
- Savitz, E. (2012). *Big data in the enterprise: A lesson or two from big brother*. Retrieved from <http://www.forbes.com/sites/ciocentral/2012/12/26/big-data-in-the-enterprise-a-lesson-or-two-from-big-brother/>.
- Scola, N. (2013). *Obama, the 'big data' president*. Retrieved from http://www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html.
- Smith, J. (2012). *White House aims to tap power of government data*. Retrieved from <https://www.yahoo.com/news/white-house-aims-tap-power-government-data-093701014.html?ref=gs>.
- Tucker, S. (2012). *Budget pressures will drive government IT change*. Retrieved from http://www.washingtonpost.com/business/capitalbusiness/budget-pressure-will-drive-government-it-change/2012/08/24/ab928a1e-e898-11e1-a3d2-2a05679928ef_story.html.
- UN Global Pulse. (2012). *Big data for development: Challenges & opportunities*. Retrieved from UN Global Pulse, Executive Office of the Secretary-General United Nations, New York, NY at <http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf>.

White House BRAIN Initiative

Gordon Alley-Young
Department of Communications and Performing Arts,
Kingsborough Community College, City University of New York, New York, NY, USA

Synonyms

Brain research through advancing innovative neurotechnologies

Introduction

The White House BRAIN Initiative (TWHBI) includes an acronym where BRAIN stands for the Brain Research Through Advancing Innovative Neurotechnologies. The goal of the initiative is to spur brain research, such as mapping the brain's circuitry, and technology that will lead to treatments and preventions for common brain disorders. President Barack Obama first announced the initiative in his February 2013 State of the Union Address (SOTUA). More than 200 leaders from universities, research institutes, national laboratories, and federal agencies were invited to attend when President Obama formally unveiled TWHBI on April 2, 2013. The Obama administration identified this initiative as one of the grand challenges of the twenty-first century. The \$100 million initiative is funded via The National Institutes of Health (NIH), the Defense Advanced Research Projects Agency (DARPA), and the National Science Foundation (NSF) with matching support for the initiative reported to come from private research institutions and foundations. TWHBI has drawn comparisons to the Human Genome Project (HGP) for the potential scientific discovery that the project is expected to yield. The HGP and TWHBI are also big data projects for the volume of data that they have already produced and will produce in the future.

History and Aims of the Initiative

TWHBI aims to provide opportunities to map, study, and thus treat brain disorders including Alzheimer's disease, epilepsy, autism, and traumatic brain injuries. The NIH will lead efforts under the initiative to map brain circuitry, measure electrical/chemical activity along those circuits, and understand the role of the brain in human behavioral and cognitive output. The initiative is guided by eight key goals. The first is to make various types of brain cells available for experimental researchers to study their role in illness and well-being. The second is to create multilayered maps of the brain's different circuitry levels as well as a map of the whole organ. The third

would see the creation of a dynamic picture of the brain through large-scale monitoring of neural activity. Fourth is to link brain activity to behavior with tools that could intervene in and change neural circuitry. A fifth goal is to increase understanding of the biological basis for mental processes by theory building and developing new data analysis tools. The sixth is to innovate technology to better understand the brain so as to better treat disorders. The seventh is to establish and sustain interconnected networks of brain research. Finally, the last goal is to integrate the outcomes of the other goals to discover how dynamic patterns of neural activity get translated into human thought, emotion, perception, and action in illness and in health.

NIH Director Dr. Francis Collins echoed President Obama in publically stating that TWHBI will change the way we treat the brain and grow the economy (National Institutes of Health 2014). During his 2013 SOTUA, President Obama drew an analogy to the Human Genome Project (HGP) arguing that for every dollar the USA invested in the project, the US economy gained \$140. Estimates suggest that the HGP created \$800 billion in economic activity. The HGP was estimated to cost \$3 billion and take 15 years (i.e., 1990–2005). The project finished 2 years early and under cost at \$2.7 billion in 1991 dollars. The HGP project is estimated to have cost \$3.39–\$5 billion in 2003 dollars. TWHBI has a budget of \$100 million allocated in budget year 2014 with comparable funds (\$122 million) contributed by private investors. A US federal report calls for \$4.5 billion in funding for brain research over the next 12 years.

Projects Undertaken by the Initiative

The first research paper believed to be produced under TWHBI initiative came from a paper published on June 19, 2014, by principal investigator Dr. Karl Deisseroth of Stanford University. The research described Deisseroth and his team's innovation of the CLARITY technique that can remove fat from the brain without damaging its wiring and enable the imaging of a whole

transparent brain. Data from the study is being used by international biomedical research projects.

TWHBI was undertaken because it addresses what the science, society, and government considers one of the grand challenges of the twenty-first century (i.e., The HGP was previously deemed a grand challenge). Unlocking the secrets of the brain will tell us how the brain can record, process, utilize, retain, and recall large amounts of information. Dr. Geoffrey Ling, deputy director of the Defense Sciences Office at Defense Advanced Research Projects Agency (DARPA), states that TWHBI is needed to attract young and intelligent people into the scientific community. Ling cites a lack of available funding as a barrier to persuading students to pursue research careers (Vallone 2013). Current NIH director and former HGP director Dr. Francis Sellers Collins notes the potential of TWHBI to create jobs while potentially curing diseases of the brain and the nervous system, for instance, Alzheimer's disease (AD). In 2012 Health and Human Services Secretary Kathleen Sebelius stated the Obama administration's goal to cure AD by 2025. The Alzheimer's Association (AA) estimates that AD/dementia health and care cost \$203 billion in 2013 (\$142 billion by Medicare/Medicaid); this will reach \$1.2 trillion by 2050 (Alzheimer's Association 2013).

Dr. Ling argues that for scientists to craft and validate their hypotheses to build on their knowledge that potentially lead to medical breakthroughs, they need access to the latest research tools. Ling states that some of the today's best clinical brain research tools are nonetheless limited and outdated in light of TWHBI work that remains to be done. To bolster his case for better research tools, Ling uses an analogy whereby the physical brain is hardware and the dynamic processes across the brain's circuits are software. Ling notes that cutting-edge tools can help identify bugs in the brain's software caused by a physical trauma (i.e., to the hardware) that once found might be repairable. The tools necessary for medical research will need to be high-speed tools with a much greater capacity for record signals from brain cells. TWHBI, by bringing together

scientists and researchers from a variety of fields such as nanoscience, imaging, engineering, informatics, has the greatest opportunity to develop these tools.

Earlier Efforts and Influences

Brain research was emphasized prior to TWHBI by the previous two administrations. The Clinton administration held a White House conference on early childhood development and leaning focused on insights gleaned from the latest brain research in 1997. In 2002 the Bush administration's National Drug Control Policy Director John Walters donated millions of dollars of drug-war money to purchase dozens of MRI machines. Their goal was a decade long, \$100 million brain-imaging initiative to study the brain to better understand addiction.

Publicity surrounding TWHBI brings attention to how much science has learned about the brain in relatively short period of time. In the nineteenth century, brain study focused mostly on what happens when parts of the brain are damaged/removed. For instance, Phineas Gage partially lost his prefrontal cortex in an 1848 accident, and scientists noted how Mr. Gage changed from easygoing and dependable before to angry and irresponsible afterward. From the late eighteenth to mid-nineteenth centuries, pseudoscientists practiced phrenology or reading a person's mind by handling a person's skull.

Phillip Low, a director of San Diego-based NeuroVigil Inc. (NVI), states that the White House talked to many scientists and researchers while planning TWHBI but did not reveal to these individuals that they were talking to many others, all of who potentially believed they were the parent of TWHBI. However, the originators of the idea that lead to TWHBI are said to be six scientists, whose journal article in the June 2012 issue of *Neuron* proposed a brain-mapping project. The six are A. Paul Alivisatos (University of California Berkeley), Miyoung Chun (The Kavli Foundation), George M. Church (Harvard University), Ralph J. Greenspan (The Kavli Institute), Michael L. Roukes (Kavli Nanoscience Institute), and

Rafael Yuste (Columbia University) (Alivisatos et al. 2012). *New York Times* reporter Steve Connor says the roots of TWHBI occur 10 years earlier when Microsoft cofounder and philanthropist Paul G. Allen established a brain science institute in Seattle for a \$300 million investment. Similarly, with a \$500 million investment, billionaire philanthropist Fred Kavli funded brain institutes at Yale, Columbia, and the University of California (Broad 2014). It was primarily scientists from these two institutes that crafted the TWHBI blueprint. Connor states that there are benefits and downsides to TWHBI's connections to private philanthropy. Connor acknowledges that philanthropists are able to invest in risky initiatives in a way that the government cannot but that this can lead to a self-serving research focus, the privileging of affluent universities at the expense of poorer ones and a US government that is following the lead of private interests rather than setting the course itself (Connor 2013).

The \$100 million for the first phase of TWHBI in fiscal year 2014 comes from three government agencies' budgets specifically NIH, DARPA, and NSF. The NIH Blueprint for Neuroscience Research will lead with contributions specifically geared to projects that would lead to the development of cutting edge, high-speed tools, training, and other resources. The next generation of tools has designated as viewed as vital to the advancement of this initiative. Contributor DARPA will invest in programs that aim to understand the dynamic functions of the brain, noted in Dr. Ling's analogy as the software of the brain, showing breakthrough applications based on the dynamic function insights gained. DARPA also seeks to develop new tools for capturing and processing dynamic neural and synaptic activities. DARPA develops applications for improving the diagnosis and treatment of post-traumatic stress, brain injury, and memory loss sustained through war and battle. Such applications would include generating new information processing systems related to the information processing system in the brain and mechanisms of functional restoration after brain injury. DARPA is mindful that advances in neurotechnology, such as those outlined above, will entail ethical, legal, and social issues that it

will oversee via its own experts. Ethics are also at the forefront of TWHBI. Specifically President Obama identified adhering to the highest standards of research protections as a prime focus. Oversight of ethical issues related to this as well as any other neuroscience initiative will fall to the administration's Commission for the Study of Bioethical Issues.

The NSF's strength as a contributor to TWHBI is that it will sponsor interdisciplinary research that spans the fields of biology, physics, engineering, computer science, social science, and behavioral science. The NSF's contribution to TWHBI again emphasizes the development of tools and equipment specifically molecular-scale probes that can sense and record the activity of neural networks. Additionally, the NSF will also seek to address the innovations that will be necessary in the field of big data in order to store, organize, and analyze the enormous amounts of data that will be produced. Finally, NSF projects under TWHBI will see better understanding of how thoughts, emotions, actions, and memories get represented in the brain.

In addition to federal government agencies, at least four private institutes and foundations have pledged an estimated \$122 million to support to TWHBI: The Allen Institute (TAI), the Howard Hughes Medical Institute (HHMI), The Kavli Foundation (TKF), and The Salk Institute for Biological Studies (TSI). TAI's strengths lie in large-scale brain research, tools, and data sharing which is necessary for a big data project like TWHBI represents. Starting in March 2012, TAI undertook a 10-year project to unlock the neural code (i.e., how brain activity leads to perception, decision-making, and action). HHMI by comparison is the largest nongovernmental funder of basic biomedical research and has long supported neuroscience research. TKF anticipates drawing on the endowments of existing Kavli Institutes (KI) to fund its participation in TWHBI. This includes funding new KIs. Finally the TSI, under its dynamic BRAIN initiative, will support cross-boundary research in neuroscience. For example, TSI researchers will map brain's neural networks to determine their interconnections. TSI scientists will lay the groundwork for solving neurological puzzles such as Alzheimer's/Parkinson's by

studying age-related brain differences (The White House 2013).

The work of TWHBI will be spread across affiliated research institutions and laboratories across the USA. The NIH is said to be establishing a bicoastal cochaired working group under Dr. Cornelia Bargmann, a former UCSF Professor, with the Rockefeller University in New York City and Dr. William Newsome from California's Stanford University to specify goals for the NIH's investment and create a multiyear plan for achieving these goals with timelines and costs (University of California San Francisco 2013). On the east coast of the USA, the NIH Blueprint for Neuroscience Research, draws on 15 of its 27 NIH Institutes and Centers headquartered in Bethesda, MD, will be a leading NIH contributor to TWHBI. Research will occur in nearby Virginia at HHMI's Janelia Farm Research Campus that focuses on developing new imaging technologies and finding out how information is stored and processed in neural networks. Imaging technology furthers TWHBI's goals of mapping the brain's structures by allowing researchers to create dynamic brain pictures down to the level of single brain cells as they interact with complex neural circuits at the speed of thought.

Conclusion

Contributions to and extensions of TWHBI are also happening on the US west coast and internationally. San Diego State University (SDSU) is contributing to TWHBI via its expertise in clinical and cognitive neuroscience specifically their investigations to understand and treat brain-based disorders like autism, aphasia, fetal alcohol spectrum (FAS) disorders, and AD. San Diego's NVI, founded in 2007 and advised by Dr. Stephen Hawking, and its founder, CEO, and Director Dr. Phillip Low, helped to shape TWHBI initiative. NVI's is notable for its iBrain™ single-channel electroencephalograph (EEG) device that noninvasively monitors the brain (Keshavan 2013). Dr. Low has also taken the message of the WBHI international as he was asked to go to Israel and help them develop their own BRAIN initiative. To this end Dr. Low delivered one of two

keynotes for Israel's first International Brain Technology Conference in Tel Aviv in October 2013. Australia also supports TWHBI through neuroscience research collaboration and increased hosting of the NSF's US research fellows for collaborating on relevant research projects.

Cross-References

- Big Data
- Data Sharing
- Medicaid

References

- Alivisatos, A. P., Chun, M., Church, G. M., Greenspan, R. J., Roukes, M. L., & Yuste, R. (2012). The brain activity map project and the challenge of functional connectomics. *Neuron*, 74(6), 970–974.
- Alzheimer's Association. (2013). *Alzheimer's Association applauds White House Brain Mapping Initiative*. Retrieved from Alzheimer's Association National Office, Chicago, IL at http://www.alz.org/news_and_events_alz_association_applauds_white_house.asp.
- Broad, W.J. (2014). *Billionaires with big ideas are privatizing American science*. Retrieved from The New York Times, New York, NY <http://www.nytimes.com/2014/03/16/science/billionaires-with-big-ideas-are-privatizing-american-science.html>.
- Connor, S. (2013). *One of the biggest mysteries in the universe is all in the head*. Retrieved from Independent Digital News and Media, London, UK at <http://www.independent.co.uk/voices/comment/one-of-the-biggest-mysteries-in-the-universe-is-all-in-the-head-8791565.html>.
- Keshavan, M. (2013). BRAIN Initiative will tap our best minds. *San Diego Business Journal*, 34(15), 1.
- National Institutes of Health. (2014). *NIH embraces bold, 12-year scientific vision for BRAIN Initiative*. Retrieved from National Institutes of Health, Bethesda, MD at <http://www.nih.gov/news/health/jun2014/od-05.htm>.
- The White House. (2013). *Fact sheet: BRAIN Initiative*. Retrieved from The White House Office of the Press Secretary, Washington, DC at <http://www.whitehouse.gov/the-press-office/2013/04/02/fact-sheet-brain-initiative>.
- University of California San Francisco. (2013). *President Obama unveils brain mapping project*. Retrieved from the University Of California San Francisco at <http://www.ucsf.edu/news/2013/04/104826/president-obama-unveils-brain-mapping-project>.
- Vallone, J. (2013). Federal initiative takes aim at treating brain disorders. In *Investors Business Daily*, Los Angeles, CA, (p. A04).

WikiLeaks

Kim Lacey

Saginaw Valley State University, University
Center, MI, USA

WikiLeaks is a nonprofit organization devoted to sharing classified, highly secretive, and otherwise controversial documents to promote transparency among global superpowers. These shared documents are commonly referred to as “leaks.” WikiLeaks has received both highly positive and negative attention for this project particularly because of its mission to share leaked information. WikiLeaks is operated by the Icelandic Sunshine Press, and Julian Assange is often named the founder of the organization.

WikiLeaks began in 2006, and its founding is largely attributed to Australian Julian Assange, often described as an Internet activist and hacker. The project, which aims to share government documents usually kept from citizens, is a major source of division between individuals and officials. The perspective on this division differs depending on the viewpoint. From the perspective of its opponents, the WikiLeaks documents are obtained illegally, and their distribution is potentially harmful for national security purposes. From the perspective of its supporters, the documents point to egregious offenses perpetrated, and ultimately stifled, by governments. On its website, WikiLeaks notes that it is working toward what it calls “open governance,” the idea that leaks are not only for international, bureaucratic diplomacy but more importantly for clarity of citizens’ consciousness.

In 2010, Chelsea (born Bradley) Manning leaked a United States’ military cable containing 400,000 files regarding the Iraq War. According to Andy Greenberg, this leak, which later became known as Cablegate, marked the largest leak of United States’ government information since Daniel Ellsberg photocopied The Pentagon Papers. After chatting for some time, Manning confessed to former hacker Adrian Lamo. Eventually, Lamo turned Manning over to the army

authorities leading to her arrest. The United States’ government officials were outraged by the leak of classified documents and viewed Manning as a traitor. This leak eventually led to Manning’s detention, and officials kept her detained for more than 1,000 days without a trial. Because of this delay, supporters of WikiLeaks were outraged at Manning’s denial of a swift trial. Manning was eventually acquitted of aiding the enemy, but, in August 2013, was sentenced to 35 years for various crimes including violations of the Espionage Act.

One of the most well-known documents Manning shared put WikiLeaks on the map for many who were previously unfamiliar with the organization. This video, known familiarly as “Collateral Murder,” shows a United States’ Apache helicopter shooting Reuters reporters, individuals helping these reporters, and seriously injuring two children. There have been two versions of the video that have been released: a shorter, 17-min video and a more detailed 39-min video. Both videos were leaked by WikiLeaks and remain on its website.

WikiLeaks uses a number of different drop boxes in order to obtain documents and maintain the anonymity of the leakers. Many leakers are well versed in anonymity protective programs such as Tor, which uses what they call “onion routing”: several layers of encryption to avoid detection. However, in order to make leaking less complicated, WikiLeaks provides instructions on its website for users to skirt around regular detection through normal identifiers. Users are instructed to submit documents in one of many anonymous drop boxes to avoid detection.

In order to verify the authenticity of a document, WikiLeaks performs several forensic tests including weighing the price of forgery as well as possible motives for falsifying information. On its website, WikiLeaks explains that it verified the now infamous “Collateral Murder” video by actually sending journalists to interview individuals affiliated with the attack. WikiLeaks states that simply when it publishes a document, the fact that it has been published is verification enough. By making information more freely available, WikiLeaks aims to start a larger conversation

within the press about access to authentic documents and democratic information.

Funding for WikiLeaks has been a contentious issue since its founding. Since 2009, Assange has noted several times that WikiLeaks is in danger of running out of funding. One of the major reasons causing these funding shortages is the result of many corporations (including Visa, MasterCard, and PayPal) ceasing to allow its customers to donate money to WikiLeaks. On the WikiLeaks website, this action is described as the “banking blockade.” To work around this banking blockade, many mirror sites (websites that are hosted separately but contain the same information) have appeared allowing users to access WikiLeaks documents and also donate with “blocked” payment methods. WikiLeaks also sells paraphernalia on its website, but it is unclear if these products fall under the banking blockade restrictions.

Because of his affiliation with WikiLeaks, Julian Assange has been granted political asylum in Ecuador in 2012. Prior to his asylum, he had been accused of molestation and rape in Sweden but evaded arrest. In June 2013, Edward Snowden, a former employer of the National Security Agency (NSA), leaked evidence of the United States spying on its citizens to the UK’s *The Guardian*. On many occasions, WikiLeaks has supported Snowden, helping him apply for political asylum, providing funding, and also providing him with escorts him on flights (most notably Sarah Harrison accompanying Snowden from Hong Kong to Russia).

WikiLeaks has been nominated for multiple awards for reporting. Among the awards, it has won including the Economist Index on Censorship Freedom of Expression award (2008) and the Amnesty International human rights reporting award (2009, New Media). In 2011, Norwegian citizen Snorre Valen publically announced that he nominated Julian Assange for the Nobel Peace Prize, although Assange did not win.

Cross-References

- [National Security Agency \(NSA\)](#)
- [Transparency](#)

Further Reading

- Dwyer, D. n.d. “WikiLeaks’ Assange for Nobel Prize?” *ABC News*. Available at: <http://abcnews.go.com/Politics/wikileaks-julian-assange-nominated-nobel-peace-prize/story?id=12825383>. Accessed 28 Aug 2014.
- Greenberg, A. *This machine kills secrets: How wikileakers, cypherpunks, and hacktivists aim to free the world’s information*. Dutton: New York, 2012.
- Sifry, Micah L 2011. *WikiLeaks and the age of transparency*. O/R Books: New York, Wikileaks.org.
- WikiLeaks*. Available at: <https://www.wikileaks.org/>. Accessed 28 Aug 2014.
- Tate, J. n.d., “Bradley Manning Sentenced to 35 Years in WikiLeaks Case.” *Washington Post* Available at: http://www.washingtonpost.com/world/national-security/judge-to-sentence-bradley-manning-today/2013/08/20/85bee184-09d0-11e3-b87c-476db8ac34cd_story.html. Accessed 26 Aug 2014.
- WikiRebels: The Documentary*. n.d.. Available at: <https://www.youtube.com/watch?v=z9xrO2Ch4Co>. Accessed 1 Sept 2012.

Wikipedia

Ryan McGrady

North Carolina State University, Raleigh, NC, USA

Wikipedia is an open-access online encyclopedia hosted and operated by the Wikimedia Foundation (WMF), a San Francisco-based nonprofit organization. Unlike traditional encyclopedias, Wikipedia is premised on an open editing model whereby everyone using the site is allowed and encouraged to contribute content and make changes. Since its launch in 2001, it has grown to over 40 million articles across nearly three hundred languages, constructed almost entirely by unpaid pseudonymous and anonymous users. Since its infancy, Wikipedia has attracted researchers from many disciplines to its vast collection of user-generated knowledge, unusual production model, active community, and open approach to data.

Wikipedia works on a type of software called a wiki, a popular kind of web application designed to facilitate collaboration. Wiki pages can be modified directly using a built-in text editor. When a

user saves his or her changes, a new version of the article is created and immediately visible to the next visitor. Part of what allows Wikipedia to maintain standards for quality is the meticulous record-keeping of changes provided by wiki software, storing each version of a page permanently in a way that is easily accessible. If someone makes changes that are not in the best interest of the encyclopedia, another user can easily see the extent of those changes and if necessary restore a previous version or make corrections. Each change is timestamped and attributed to either a username or, if made anonymously, an IP address. Although Wikipedia is transparent about what data it saves and draws little criticism on privacy matters, any use of a wiki requires self-awareness given that one's actions will be archived indefinitely.

Article histories largely comprise the Wikipedia database, which the WMF makes available to download for any purpose compatible with its Creative Commons license, including mirroring, personal and institutional offline use, and data mining. The full English language database download amounts to more than ten terabytes, with several smaller subsets available that, for example, exclude discussion pages and user profiles or only include the most current version of each page.

As with any big data project, there is a challenge in determining not just what questions to ask but how to use the data to convey meaningful answers. Wikipedia presents an incredible amount of knowledge and information, but it is widely dispersed and collected in a database organized around articles and users, not structured data. One way the text archive is rendered intelligible is through visualization, wrangling the unwieldy information by expressing statistics and patterns through visuals like graphs, charts, or histograms. Given the multi-language and international nature of Wikipedia, as well as the disproportionate size and activity of the English version in particular, geography is important in its critical discourse. Maps are thus popular visuals to demonstrate disparities, locate concentrations, and measure coverage or influence. Several programs have been developed to create visualizations using

Wikipedia data, as well. One of the earliest, the IBM History Flow tool, produces images based on stages of an individual article's development over time, giving a manageable, visual form to an imposingly long edit history and the disagreements, vandalism, and controversies it contains.

The Wikipedia database has been and continues to be a valuable resource, but there are limitations to what can be done with its unstructured data. It is downloaded as a relational database filled with text and markup, but machines that researchers use to process data are not able to understand text like a human, limiting what tasks they can be given. It is for this reason there have been a number of attempts to extract structured data as well. DBPedia is a database project started in 2007 to put as much of Wikipedia into the Resource Description Framework (RDF) as possible. Whereas content on the web typically employs HTML to display and format text, multimedia, and links, RDF emphasizes not what a document looks like but how its information is organized, allowing for arbitrary statements and associations which effectively make the items meaningful to machines. The article for the film *Moonlight Kingdom* may contain the textual statement "it was shot in Rhode Island," but a machine would have a difficult time extracting the desired meaning, instead preferring to see a subject "Moonlight Kingdom" with a standard property "filming location" set to the value "Rhode Island."

In 2012, WMF launched Wikidata, its own structured database. In addition to Wikipedia, WMF operates a number of other sites like Wiktionary, Wikinews, Wikispecies, and Wikibooks. Like Wikipedia, these sites are available in many languages, each more or less independent from the others. To solve redundancy issues and to promote resource sharing, the Wikimedia Commons was introduced in 2004 as a central location for images and other media for all WMF projects. Wikidata works on a similar premise with data. Its initial task was to centralize inter-wiki links, which connect, for example, the English article "Cat" to the Portuguese "Gato" and Swedish "Katt." Inter-language links had previously been handled locally, creating links at the

bottom of an article to its counterparts at every other applicable version. Since someone adding links to the Tagalog Wikipedia is not likely to speak Swedish, and because someone who speaks Swedish is not likely to actively edit the Tagalog Wikipedia and vice versa, this process frequently resulted in inaccurate translations, broken links, one-way connections, and other complications. Wikidata helps by acting as a single junction for each topic.

A topic, or an item, on Wikidata is given its own page which includes an identification number. Users can then add a list of alternative terms for the same item and a brief description in every language. Items also receive statements connecting values and properties. For example, The Beatles's 1964 album *A Hard Day's Night* is item Q182518. The item links to the album's Wikipedia articles in 49 languages and includes 17 statements with properties and values. The very common *instance of* property has the value "album," a property called *record label* has the value "Parlophone Records," and four statements connect the property *genre* with "rock and roll," "beat music," "pop music," and "rock music." Other statements describe its recording location, personnel, language, and chronology, and many applicable properties are not yet filled in. Like Wikipedia, Wikidata is an open community project and anybody can create or modify statements. Some of the other properties items are given include names, stage names, pen names, dates, birth dates, death dates, demographics, genders, professions, geographic coordinates, addresses, manufacturers, alma maters, spouses, running mates, predecessors, affiliations, capitals, awards won, executives, parent companies, taxonomic orders, and architects, among many others. So as to operate according to the core Wikipedia tenet of neutrality, multiple conflicting values are allowed. Property-value pairs can furthermore be assigned their own property-value pairs such that the *record sales* property and its value can have the qualifier *as of* and another value to reflect when the sales figure was accurate. Each property-value pair along the way can be assigned references akin to cited sources on Wikipedia.

Some Wikipedia metadata is easy to locate and parse as fundamental elements of wiki technology: timestamps, usernames, and article titles, for example. Other data is incidental, like template parameters. Design elements that would otherwise be repeated in many articles are frequently copied into a separate template which can then be invoked when relevant, using parameters to customize it for the particular page on which it is displayed. For example, in the top-right corner of articles about books there is typically a neatly formatted table called an infobox which includes standardized information input as template parameters like author, illustrator, translator, awards received, number of pages, Dewey decimal classification, and ISBN number. A fundamental part of DBPedia and the second goal for Wikidata is the collection of data based on these relatively few structured fields that exist in Wikipedia.

Standardizing the factual information in Wikipedia holds incredible potential for research. Wikidata and DBPedia, used in conjunction with the Wikipedia database, make it possible to, for example, assess article coverage of female musicians as compared to male musicians in different parts of the world. Since they use machine-readable formats, they can also interface with one another and with many other sources like GeoNames, Library of Congress Subject Headings, Internet Movie Database, MusicBrainz, and Freebase, allowing for richer, more complex queries. Likewise, just as these can be used to support Wikipedia research, Wikipedia can be used to support other forms of research and even enhance commercial products. Google, Facebook, IBM, and many others regularly make use of data from Wikipedia and Wikidata in order to improve search results or provide better answers to questions. By creating points of informational intersection and interpretation for hundreds of languages, Wikidata also has potential for use in translation applications and to enhance cultural education. The introduction of Wikidata in 2012, built on an already impressively large knowledge base, and its ongoing development, have opened many new areas for exploration and accelerated the pace of experimentation, incorporating the

data into many areas of industry, research, education, and entertainment.

Cross-References

- [Anonymity](#)
- [Crowdsourcing](#)
- [Open Data](#)

Further Reading

- Jemielniak, D. (2014). *Common knowledge: An ethnography of wikipedia*. Stanford: Stanford University Press.
- Kröttscha, M., et al. (2007). Semantic Wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4), 251–261.
- Leetaru, K. (2012). A bigdata approach to the humanities, arts, and social sciences: Wikipedia's view of the world through supercomputing. *Research Trends*, 30, 17–30.
- Stefaner, M., et al. Notability – Visualizing deletion discussions on Wikipedia. <http://www.notabilia.net/>.
- Viégas, F., et al. (2004). *Studying cooperation and conflict between authors with history flow visualizations*. Paper presented at CHI 2004, Vienna.

natural resource management. In addition to the financial support, the World Bank provides policy advice, research, analysis, and technical assistance to various countries in order to inform its own investments and ultimately to work toward its key objectives. Part of its activities relate to the provision of tools to research and address development challenges, some of which are in the form of providing access to data, for example, the Open Data website which includes a comprehensive range of downloadable data sets related to different issues. This shows its recognition of the demand for access to quantitative data to inform development strategies (Lehdonvirta and Ernkvist 2011).

A significant amount of the data hosted and disseminated by the World Bank is drawn from national statistical organizations, and it recognizes that the quality of global data therefore is reliant on the capacity and effectiveness of these national statistical organizations. The World Bank has ten key principles with respect to its statistical activities (in line with the Fundamental Principles of Official Statistics and the Principles Governing International Statistical Activities of the United Nations Statistical Division): quality, innovation, professional integrity, partnership, country ownership, client focus, results, fiscal responsibility, openness, and good management.

The world is now experiencing unprecedented capacity to generate, store, process, and interact with data (McAfee and Brynjolfsson 2012), a phenomenon that has been recognized by the World Bank, like other international institutions. For the World Bank, data is seen as critical for the design, implementation, and evaluation of efficient and effective development policy recommendations. In 2014, Jim Yong Kim, the President of the World Bank, discussed the importance of efforts to invest in infrastructure, including data systems. Big data is recognized as a new advancement which has the potential to enhance efforts to address development, although it recognizes there are a series of challenges associated with this. In 2013, the World Bank hosted an event where over 150 experts, data scientists, civil society groups, and development practitioners met to analyze various forms of big data and consider

World Bank

Jennifer Ferreira
Centre for Business in Society, Coventry
University, Coventry, UK

The World Bank, part of the World Bank Group established in 1944, is the international financial institution responsible for promoting economic development and reducing poverty. The World Bank has two key objectives: to end extreme poverty by reducing the proportion of the world's population living on less than \$1.25 a day and promoting shared prosperity by fostering income growth in the lowest 40% of the population.

A core activity for the World Bank is the provision of low interest loans, zero- to low-interest grants to developing countries. This could be to support a wide range of activities from education and health care to infrastructure, agriculture, or

how it could be used to tackle development issues. The event was a public acknowledgement of how the World Bank viewed the importance of expanding the awareness of how big data can help combine various data sets to generate knowledge which can in turn foster development solutions.

A report produced in conjunction with the World Bank, *Big Data in Action for Development*, highlights some of the potential ways in which big data can be used to work toward development objectives and some of the challenges associated with doing so. The report sets out a conceptual framework for using big data in the development sector highlighting the potential transformative capacity of big data, particularly in relation to raising awareness, developing understanding, and contributing to forecasting.

Using big data to develop and enhance awareness of different issues has been widely acknowledged. Examples of this include: using demographic data in Afghanistan to detect impacts of small scale violence outbreaks, using social media content to indicate unemployment rises or crisis related stress, or using tweets to recognize where cholera outbreaks were appearing at a much faster rate than was recognized in official statistics. This ability to gain awareness of situations, experiences, and sentiments is seen to have the potential to reduce reaction times and improve processes which deal with such situations.

Big data can also be used to develop understanding of societal behaviors (LaValle et al. 2011). Examples include investigation of twitter data to explore the relationship between food and fuel price tweets and changes in official price indexes in Indonesia; after the 2010 earthquake in Haiti, mobile photo data was used to track population displacement after the event, and satellite rainfall data was used in combination with qualitative data sources to understand how rainfall affects migration.

Big data is also seen to have potential for contributing to modelling and forecasting. Examples include: the use of GPS-equipped vehicles in Stockholm, providing real-time traffic assessments, which are used in conjunction with other

data sets such as weather which can then be used to make traffic predictions, using mobile phone data to predict mobility patterns.

The World Bank piloted some activities in Central America to explore the potential of big data to impact on development agendas. This region has historically experienced low frequencies of data collection for traditional data forms, such as household surveys and so other forms of data collection, were viewed as particularly important. One of these pilot studies used google trends data to explore the potential for the ability to forecast price changes to commodities. Another study, in conjunction with the UN Global Pulse, explored the use of social media content to analyze public perceptions of policy reforms, in particular a gas subsidy reform in El Salvador, highlighting the potential for this form of data to complement other studies on public perception (United Nations Global Pulse 2012).

The report from the World Bank, *Big Data in Action for Development*, presents a matrix of different ways in which big data could be used in transformational ways toward the development agenda: using mobile data (e.g., reduced mobile phone top ups as an indicator of financial stress), financial data (e.g., increased understanding of customer preferences), satellite data (e.g., to crowd source information on damage after an earthquake), internet data (e.g., to collect daily prices), and social media data (e.g., to track parents perception of vaccination). The example of examining the relationship between food and fuel prices and corresponding change in official price index measures by using twitter data (by the UN Global Pulse Lab) is outlined in detail explaining how it was used to provide an indication of social/economic conditions in Indonesia. This was done by extracting tweets mentioning food and fuel prices between 2011 and 2013 (around 100,000 relevant tweets after filtering for location and language) and analyzing these with corresponding changes from official data sets. The analysis indicated a clear relationship between official food inflation statistics and the number of tweets about food price increases. This study was cited as an example of how big data could be used to analyze public sentiment, in addition to objective

economic conditions. The examples mentioned here are just some of the activities undertaken by the World Bank to embrace the world of big data.

As with many other international institutions which recognize the potential uses for big data, the World Bank also recognizes there are a range of challenges associated with the generation, analysis, and use of big data.

One of the most basic challenges for many organizations (and individuals) is gaining access to data, from both government institutions and the private sector. A new ecosystem needs to be developed where data is made openly available and sharing incentives are in place. It is acknowledged by the World Bank that international agencies will need to address this challenge by not only by promoting the availability of data but promoting collaboration and mechanisms for sharing data. In particular, a shift in business models will be required in order to ensure the private sector is willing to share data, and governments will need to design policy mechanisms to ensure the value of big data is captured and is shared across departments. Related to this, there need to be considerations of how to engage the public with this data.

Thinking particularly about the development agenda at the heart of the World Bank, there is a paradox: countries where poverty is high or where development agendas require the most attention are often countries where data infrastructures or technological systems are insufficient. Because the generation of big data relies largely on technological capabilities, relying on those who use or interact with digital sources may be systematically unrepresentative of the larger population that forms the focus of the research.

The ways in which data are recorded have implications for the results which are interpreted. Where data is passively recorded, there is less potential for bias in the results generated, and likewise where data is actively recorded, there is greater potential for the results to be more susceptible to selection bias. Furthermore, how data is processed into a more structured form from the often very large and unstructured data sets requires expertise to both clean the data and where necessary aggregate it (e.g., if one set of data collected every hour, and another every day).

Then the media through which data is collected is also an important factor to consider. Mobile phones, for example, producing highly sensitive data, satellite images produce highly unstructured data, and social media platforms produce a lot of unstructured text which requires filtering and codifying which in itself requires specific analytic capabilities.

Then in order to make effective use of big data, those using it need to consider elements about the data itself. The generation of big data has been driven by advances in technology, yet these advances are not alone sufficient to be able to understand the results which can be gleaned from big data. Transforming vast data sets into meaningful results requires effective human capabilities. Depending on how the data is generated, and by whom, there is scope for bias and therefore misleading conclusions. Then with large amounts of data, there is a tendency for patterns to be observed where there may be none; because of its nature, big data can give rise to significant statistical correlations. It is important to remember that correlation does not imply causation. Then just because there is large amount of data available, this does not necessarily mean this is the right data for the question or issue being investigated.

The World Bank acknowledges that for big data to be made effective for development, there will need to be collaboration between practitioners, social scientists, and data scientists in order to ensure the understanding of the real-world conditions and data generation mechanisms, and methods of interpretation are effectively combined. Beyond this there will need to be cooperation between public and private sector bodies in order to foster greater data sharing and incentivize the use of big data across different sectors. Even when data has been accessed, in nearly all occasions it needs to be filtered and made suitable for analysis. Filters require human input and need to be applied carefully as their use may preclude information and affect the results. Data needs to be cleaned. Mobile data is received in unstructured form of millions of files, which requiring time-intensive processing to obtain data suitable for analysis. Likewise, analysis of text

from social media requires a decision making process to filter out suitable search terms.

Finally, there are a series of concerns about how privacy is ensured with big data, given that often there are elements of big data which can be sensitive in nature (either to the individual or commercially). This is made more complicated as each country will have different regulations about data privacy which poses particular challenges for institutions working across national boundaries, like the World Bank.

For the World Bank, the use of big data is seen to have potential for improving and changing the international development sector. Underpinning the ideas of the World Bank's approach to big data is the recognition that while the technological capacities for generation, storage, and processing of data continue to develop, this also needs to be accompanied by institutional capabilities to enable big data analysis to contribute to effective actions that can contribute to development, whether this is through strengthening of warning systems, raising awareness, or developing understanding of social systems or behaviors.

The World Bank has begun to consider an underlying conceptual framework around the use of big data, in particular considering the challenges it presents in terms of using big data for development. In the report *Big Data in Action for Development*, it is acknowledged that there is great potential for big data to provide a valuable input for designing effective development policy recommendation but also that big data is no panacea (Coppola et al. 2014). The World Bank has

made clear efforts to engage with the use of big data and has begun to explore areas of clear potential for big data use. However, questions remain about how it can support countries to take ownership and create, manage, and maintain their own data, contributing to their own development agendas in effective ways.

Cross-References

- [International Development](#)
- [United Nations Educational, Scientific and Cultural Organization \(UNESCO\)](#)

Further Reading

- Coppola, A., Calvo-Gonzalez, O., Sabet, E., Arjomand, N., Siegel, R., Freeman, C., Massarat, N. (2014). *Big data in action for development*. Washington, DC: World Bank and Second Muse. Available at: http://live.worldbank.org/sites/default/files/Big%20Data%20for%20Development%20Report_final%20version.pdf.
- LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52(2), 21–31.
- Lehdonvirta, V., & Ernkivist, M. (2011). Converting the virtual economy into development potential: Knowledge map of the virtual economy. *InfoDev/World Bank White Paper*; 1, 5–17.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 60–66.
- United Nations Global Pulse. (2012). *Big data for development: Challenges & opportunities*. New York: UN, New York.

Z

Zappos

Jennifer J. Summary-Smith
Florida SouthWestern State College, Fort Myers,
FL, USA
Culver-Stockton College, Canton, MO, USA

As one of the largest online retailers of shoes, Zappos (derived from the Spanish word *zapatos* meaning shoes) is a company that is setting an innovative trend in customer service and management style. According to Zappos' website, one of its primary goals is to provide the best online service. The company envisions a world where online customers will make 30% of all retail transactions in the United States. Zappos hopes to be the company that leads the market in online sales, setting itself aside from other online retail competitors by offering the best customer service and selection.

History of the Company

Zappos was founded in 1999 by Nick Swinmurn who developed the idea for the company while walking around a mall in San Francisco, California, looking for a pair of shoes. After spending an hour in the mall searching from store to store for the right color and shoe size, he left the mall empty handed and frustrated. Upon arriving home, Swinmurn turned to the Internet to

continue his search for his preferred shoes, which again was unsuccessful. Swinmurn realized that there were no major online retailers specializing in shoes. It was at this point that Swinmurn decided to quit his full-time job and start an online shoe retailer named Zappos. Overtime the company has evolved, focusing on making the speed of its customers' online purchase central to its business model. In order to achieve this, Zappos warehouses have everything it sells. As the company grew, it reached new heights in 2009 when Zappos and Amazon joined forces combining their passion for strong customer service. Since then, Zappos has grown significantly and restructured into ten separate companies.

Security Breach

Unfortunately, Zappos has not been without a few missteps. In 2012, the company experienced a security breach, compromising as many as 24 million customers. Ellen Messmer reports that cyberhackers successfully gained access to the company's internal network and systems. To address this security breach, Zappos CEO Tony Hsieh announced that existing customer passwords would be terminated as a result of the breach. Still yet, the cyberhackers likely gained accessed to names, phone numbers, the last four digits of credit card numbers, cryptographically scrambled passwords, email, billing information, and shipping addresses. After Zappos CEO Tony

Hsieh posted an open letter explaining the breach and how the company would head off resulting problems, there were mixed responses to how the company had handled the situation. As part of its response to the breach, the company sent out emails informing its customers of the problem urging them to change their passwords. Zappos also provided an 800-number phone service to its customers helping them through the process of choosing a new password.

However, some experts familiar with the online industry have criticized the moves by Zappos. In an article by Ellen Messmer, she interviewed an Assistant Professor of Information Technology from the University of Notre Dame, who argued that the response strategy by Zappos was not appropriate. Professor John D'Arcy posits that the company's decision to terminate customers' passwords promotes a panic mode, creating a sense of panic in its customers. In contrast, other analysts claim that Zappos public response to the situation was the right move, communicating to its customers publicly.

Nevertheless, Zappos is doing a good job of getting the information out about the security breach to the public as soon as possible, according to Professor John D'Arcy. This typically benefits the customers, creating favorable reactions. In terms of the cost of the security breaches, the Ponemon Institute estimates that on average, a data breach costs \$277 per compromised record.

Lawsuits

After the security breach, dozens of lawsuits were filed. Zappos attempted to send the lawsuits to arbitration, citing its user agreement. In the fall of 2012, a federal court struck down Zappos.com's user agreement, according to Eric Goldman. Eric Goldman is a professor of law at Santa Clara University School of Law who writes about Internet law, intellectual property, and advertising law. He states that Zappos made mistakes that are easily avoidable. The courts typically divide user agreements into one of three groups: "clickwraps" or "click-through agreements," "browsewraps," and

"clearly not a contract." Eric Goldman argues that the click-through agreements are effective in courts unlike browsewraps. Browsewraps are user agreements that bind users simply for browsing the website. The courts ruled that Zappos presented its user agreement as a browsewrap. Furthermore, Zappos claimed on its website that the company reserved the right to amend the contract whenever it saw fit. Despite other companies using this language online, it is detrimental to a contract. The courts ruled that Zappos can amend the terms of the user agreement at any time, making the arbitration clause susceptible to change as well. This makes the clause unenforceable. Eric Goldman posits that the court ruling left Zappos in a bad position because all of the risk management provisions are ineffective. In other words, losing the contract left Zappos without the following: its waiver of consequential damages, its disclaimer of warranties, its clause restricting class actions in arbitration, and its reduced statute of limitations. Conversely, companies that use click-through agreements and remove clauses that state they can amend the contract unilaterally are in a better legal position, according to Eric Goldman.

Holacracy

Zappos CEO Tony Hsieh announced in November 2013 that his company would be implementing the management style known as Holacracy. With Holacracy, there are two key elements that Zappos will follow: distributed authority and self-organization. According to an article by Nicole Leinbach-Reyhle, distribution authority allows employees to evolve the organization's structure by responding to real-world circumstances. In regard to self-organization, employees have the authority to engage in useful action to express their purpose as long as it does not "violate of the domain of another role." There is a common misunderstanding that Holacracy is nonhierarchical when in fact it is strongly hierarchical, distributing power within the organization. This approach to management creates an atmosphere where employees can speak up evolving

into leaders rather than followers. Zappos CEO Tony Hsieh states that he is trying to structure Zappos less like a bureaucratic corporation and more like a city, resulting in increased productivity and innovation. To date, with 1,500 employees, Zappos is the largest company to adopt the management model – Holacracy.

Innovation

The work environment at Zappos has become known for its unique corporate culture, which incorporates fun and humor into daily work. As stated on [Zappos.com](http://www.zappos.com), the company has a total of ten core values: “deliver WOW through service, embrace and drive change, create fun and a little weirdness, be adventurous, creative, and open-minded, pursue growth and learning, build open and honest relationships with communication, build a positive team and family spirit, do more with less, be passionate and determined, and be humble.” Nicole Leinbach-Reyhle writes that Zappos’ values help to encourage its employees to think outside of the box.

To date, Zappos is a billion-dollar online retailer, expanding beyond selling shoes. The company is also making waves in its corporate culture and hierarchy. Additionally, information technology plays a huge role in the corporation, serving its customers and the business. Based upon the growing success of Zappos, it is keeping true to its mission statement “to provide the best customer service possible.” It is evident that Zappos will continue to make positive changes for the corporation and its corporate headquarters in Las Vegas. In 2013, Zappos CEO Tony Hsieh committed \$350 million to rebuild and renovate the downtown Las Vegas region. As Sara Corbett notes in her article, he hopes to change the area into a start-up fantasyland.

Cross-References

► [Ethical and Legal Issues](#)

Further Reading

- Corbett, S. (n.d.). How Zappos’ CEO turned Las Vegas into a startup fantasyland. <http://www.wired.com/2014/01/zappos-tony-hsieh-las-vegas/>.
- Goldman, E. (n.d.). How Zappos’ user agreement Failed in court and left Zappos legally naked. <http://www.forbes.com/sites/ericgoldman/2012/10/10/how-zappos-user-agreement-failed-in-court-and-left-zappos-legally-naked/>. Accessed Jul 2014.
- Leinbach-Reyhle, N. (n.d.). Shedding hierarchy: Could Zappos be setting an innovative trend? <http://www.forbes.com/sites/nicoleleinbachreyhle/2014/07/15/shedding-hierarchy-could-zappos-be-setting-an-innovative-trend/>. Accessed Jul 2014.
- Messmer, E. (n.d.). Zappos data breach response a good idea or just panic mode? Online shoe and clothing retailer Zappos has taken assertive steps after breach, but is it enough? <http://www.networkworld.com/article/2184860/malware-cybercrime/zappos-data-breach-response-a-good-idea-or-just-panic-mode-.html>. Accessed Jul 2014.
- Ponemon Group. (n.d.). 2013 cost of data breach study: Global analysis. <http://www.ponemon.org>. Accessed Jul 2014.
- Zappos. (n.d.). <http://www.zappos.com>. Accessed Jul 2014.

Zillow

Matthew Pittman and Kim Sheehan
School of Journalism & Communication,
University of Oregon, Eugene, OR, USA

Overview and Business Model

Like most industries, real estate is undergoing dynamic shifts in the age of big data.

Real estate information, once in the hands of a few agents or title companies, is being democratized for any and all interested consumers. What were previously physical necessities – real estate agents, showings, and physical homes – are being obsolesced by digital platforms like Zillow. Real estate developers can use technology to track how communities flow and interact with one another, which will help build smarter, more efficient neighborhoods in the future. The companies that succeed in the future will be the ones who, like

Zillow, find innovative, practical, and valuable ways to navigate and harness the massive amounts of data that are being produced in and around their field.

Founded in Seattle in 2005, Zillow is a billion-dollar real estate database that uses big data to help consumers learn about home prices, rent rates, market trends, and more. They provide estimates for most housing units in the United States. It acquired its closest competitor, Trulia, in 2014 for \$3.5 billion. It is the most-viewed real estate destination in the country. Now with Trulia, it accounts for 48% of Web traffic for real estate listings, though that number is diminished to around 15% if you factor in individual realtor sites and local MLS (multiple listing service) listings. The company's chief economist Stan Humphries created a tool that processes 1.2 million proprietary statistical models three times per week on the county and state real estate data it is constantly gathering. In 2011, they shifted from an in-house computer cluster to renting space in the Amazon cloud to help with the massive computing load.

On the consumer side, Zillow is a web site or mobile app that is free to use. Users can enter a city or zip code and search, filtering out home types, sizes, or prices that are undesirable. There are options to see current homes for sale, recently sold properties, foreclosures, rental properties, and even Zillow "zestimates" (the company's signature feature) of the home's current value based on similar homes in the area, square footage, amenities, and more. Upon clicking on a house of interest, the user can see a real estate agent's description of the home, how long it has been on the market – along with any price fluctuations – as well as photos, similarly priced nearby houses, proposed mortgage rates on the home, the agents associated with it, the home's sale history, and facts and features.

Zillow makes money on real estate firms and agents that advertise through the site and by providing subscriptions to real estate professionals. They can charge more for ads that appear during a search for homes in Beverly Hills than in Bismarck, South Dakota. Some 57,000 agents spend

an average of \$4,000 every year for leads to get new buyers and sellers. Zillow keeps a record of how many times a listing has been viewed, which may help negotiate the price with among agents, buyers, and sellers. Real estate agents can subscribe to silver, gold, or platinum programs to get CRM (customer relationship management) tools, their photo in listings, a web site, and more. Basic plans start at 10 dollars a month.

Zillow's mortgage marketplace also earns them revenue. Potential homebuyers can find and engage with mortgage brokers and firms. The mortgage marketplace tells potential buyers what their monthly payment would be, how much they can afford, submit loan requests, and get quotes from various lenders. In the third quarter of 2013, Zillow's mortgage marketplace received 5.9 million loan requests from borrowers (more than all of 2011), which grew its revenue stream 120% to \$5.7 million. A majority of Zillow's revenue comes from the real estate segment that lets users browse homes for sale and for rent. This earned them over \$35 million in 2013's third quarter.

Analysts and shareholders have voiced some concerns over Zillow's business model. Zillow now spends over 70% of its revenues on sales and marketing, as opposed to 33% for LinkedIn and between 21% and 23% for IBM and Microsoft. Spending money on television commercials and online ads for its services seems to have diminishing returns for Zillow, who is spending more and more on marketing for the same net profit. What once seemed like a sure-fire endeavor – making money by connecting customers to agents through relevant and concise management of huge amounts of data – is no longer a sure thing. Zillow will have to continually evolve its business model if it is to stay afloat.

Zillow and the Real Estate Industry

Zillow has transformed the real estate industry by finding new and practical ways to make huge amounts of data accessible to common people.

Potential buyers no longer need to contact a real estate agent before searching for homes – they can start a detailed search on just about any house in the country from their own mobile or desktop device. This is empowering for consumers, but it shakes up an industry that has long relied on human agents. These agents made it their business to know specific areas, learn the ins and outs of a given community, and then help connect interested buyers to the right home. Sites that give users a tool peer into huge amounts of data (like Zillow) are useful to a point, but some critics feel only a human being who is local and present in a community can really serve potential buyers.

Because it takes an aggregate of multiple national and MLS listing sites, Zillow is rarely perfect. Any big data computing service that works with offline or subjective entities – and real estate prices certainly fit this description – will have to make logical (some would say illogical) leaps where information is scarce. When Zillow does not have exact or current data on a house or neighborhood, it “guesses” when prices come in too high, sellers have unrealistic expectations of the potential price of their home. Buyers, too, may end up paying for a home than it is actually worth.

A human expert (real estate agent) has traditionally been the expert in this area, yet people are still surprised when too much stock is put into an algorithm. Zillow zestimates tend to work best for midrange homes in an area where there are plenty of comparable houses. Zestimates are less accurate for low- and high-end homes because there are fewer comps (comparable houses for sale or recently sold). Similarly, zestimates of rural, unique, or fixer-upper homes are difficult to gauge. Local MLS sites may have more detail on a specific area, but Zillow has broader, more general information over a larger area. They estimate their coverage of American homes to be around 57%.

Real estate data is more difficult to come by in some areas. Texas doesn't provide public records of housing transaction prices, so Zillow had to access sales data from property databases through

real estate brokers. Because of the high number of cooperative buildings, New York City is another difficult area in which to gauge real estate prices. Tax assessments are made on the co-ops, not the individual units, which negates that factor in zestimate calculations. Additional information, like square footage or amenities, is also difficult to come by, forcing Zillow to seek out alternative sources.

Of course, zestimates can be accurate as well. As previously noted, when the house is midrange and in a neighborhood with plenty of comps (and thus plenty of data), zestimates can be very good indicators of the home's actual worth. As Zillow zestimates – and sources from which to draw factoring information – continue to evolve, the service may continue growing in popularity. The more popular Zillow becomes, the more incentive real estate agents will have to list all of their housing database information with the service. Agents know that, in a digital society, speed is key: 74% of buyers and 76% of sellers will work with the first agent with whom they talk.

Recently Zillow is recognizing a big shift to mobile: about 70% of Zillow's usage now occurs on mobile platforms. This trend is concurrent with other platforms' shift to mobile usage; Facebook, Instagram, Zynga, and others have begun to recognize and monetize users' access from smartphones and tablets. For real estate, this mobile activity is about more than just convenience: user can find information on homes in real time as they drive around a neighborhood, looking directly at the potential homes, and contact the relevant agent before they get home. This sort of activity bridges the traditional brick-and-mortar house hunting of the past with the instant big data access of the future (and increasingly, the present). Zillow has emerged as a leader in its field of real estate by connecting its customers, not just to big data but the right data at the right time and places.

Cross-References

► [E-Commerce](#)

Further Reading

- Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49, 45–53.
- Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N.M. (2012). The livelihoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*.
- Hagerty, J. R.(2007). How good are Zillow's estimates? *Wall Street Journal*.
- Huang, H., & Tang, Y. (2012). Residential land use regulation and the US housing price cycle between 2000 and 2009. *Journal of Urban Economics*, 71(1), 93–99.
- Wheatley, M. (n.d.). Zillow-Trulia merger will create boundless new big data opportunities. <http://siliconangle.com/blog/2014/07/31/zillow-trulia-merger-will-create-boundless-new-big-data-opportunities/>. Accessed on Sept 2014.